DOI:
10.1038/s41588-020-0599-0

Document Version
Peer reviewed version

Meta-analysis of 542,934 subjects of European ancestry identifies new genes and mechanisms predisposing to refractive error and myopia

Pirro G. Hysi[1,2,3 † *], Hélène Choquet[4†], Anthony P. Khawaja [5,6†], Robert Wojciechowski[7,8†], Milly S Tedja[9,10†], Jie Yin[4], Mark J. Simcoe[2], Karina Patasova[1], Omar A. Mahroo[1,5], Khanh K Thai[4], Phillippa M Cumberland[3,12], Ronald B. Melles[13], Virginie J.M. Verhoeven[9,10,11], Veronique Vitart[14], Ayellet Segre[15], Richard A. Stone[16], Nick Wareham[6], Alex W Hewitt[17], David A Mackey[17,18], Caroline CW Klaver[9,10,19,20], Stuart MacGregor[21], The Consortium for Refractive Error and Myopia[22], Peng T. Khaw[5], Paul J. Foster[5,23], The UK Eye and Vision Consortium[22], Jeremy A. Guggenheim[24], 23andMe Inc.[22], Jugnoo S Rahi[3,5,12,25*], Eric Jorgenson[4*], and Christopher J Hammond[1,2 *]

Author Information:

1 King's College London, Section of Ophthalmology, School of Life Course Sciences, London, UK
2 King's College London, Department of Twin Research and Genetic Epidemiology, London, UK
3 University College London, GOSH Institute of Child Health, London, UK
4 Division of Research, Kaiser Permanente Northern California, Oakland, California, USA
5 NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK
6 Department of Public Health and Primary Care, Institute of Public Health, University of Cambridge School of Clinical Medicine, Cambridge, UK
7 Department of Biophysics, Johns Hopkins University, Baltimore, MD, USA
8 Wilmer Eye Institute, Johns Hopkins School of Medicine, Baltimore, MD, USA
9 Department of Ophthalmology, Erasmus Medical Center, Rotterdam, The Netherlands
10 Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands
11 Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands
12 Ulverscroft Vision Research Group, UCL Great Ormond Street Institute of Child Health, University College London, UK
13 Kaiser Permanente Northern California, Department of Ophthalmology, Redwood City, CA, USA
14 MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, The University of Edinburgh, United Kingdom
15 Department of Ophthalmology, Harvard Medical School, Massachusetts Eye and Ear, Boston, MA, USA
16 Department of Ophthalmology, University of Pennsylvania Perelman School of Medicine, Philadelphia, Pennsylvania, USA
17 Department of Ophthalmology, Royal Hobart Hospital, Hobart, Tasmania
18 Centre for Ophthalmology and Visual Science, University of Western Australia, Lions Eye Institute, Perth, WA, WA 6009, Australia
19 Department of Ophthalmology, Radboud University Medical Center, Rotterdam
20 Institute of Molecular and Clinical Ophthalmology Basel, Switzerland
21 QIMR Berghofer Medical Research Institute, Brisbane, Australia
22 Names and affiliations of the consortium members are listed in the Supplementary Note
23 Division of Genetics and Epidemiology, UCL Institute of Ophthalmology, London, UK.
24 Cardiff University, School of Optometry & Vision Sciences, UK
25 Department of Ophthalmology and NIHR Biomedical Research Centre Great Ormond Street Hospital NHS Foundation Trust

† These authors jointly led this work
** These authors jointly supervised this work*
** email pirro.hysi@kcl.ac.uk*

Abstract

Refractive errors, in particular myopia, are a leading cause of morbidity and disability world-wide. Genetic investigation can improve understanding of the molecular mechanisms underlying abnormal eye development and impaired vision. We conducted a meta-analysis of genome-wide association studies involving 542,934 European participants and identified 336 novel genetic loci associated with refractive error. Collectively, all associated genetic variants explain 18.4% of heritability and improve the accuracy of myopia prediction (AUC=0.75). Our results suggest that refractive error is genetically heterogeneous, driven by genes participating in the development of every anatomical component of the eye. In addition, our analyses suggest that genetic factors controlling circadian rhythm and pigmentation are also involved in the development of myopia and refractive error. These results may make possible predicting refractive error and the development of personalized myopia prevention strategies in the future.

Introduction

Refractive errors occur when converging light rays from an image do not clearly focus on the retina. They are the seventh most prevalent clinical condition[1] and the second leading cause of disability in the world[2]. The prevalence of refractive error is rapidly increasing, mostly driven by a dramatic rise in the prevalence of one of its forms, myopia (near-sightedness). Although the causes of such a rise over a short time are likely due to environmental and cultural changes from the mid-20th century[3], refractive errors are highly heritable[4]. Several studies[5,6] have previously sought to identify genes controlling molecular mechanisms leading to refractive error and myopia. However, the variance and heritability that can be attributed to known genetic factors is modest[7] and our knowledge of pathogenic mechanisms remains partial. Here, we conduct a meta-analysis combining data from quantitative spherical equivalent and myopia status from large and previously unpublished genome-wide association studies (GWAS) of more than half a million subjects from the UK Biobank, 23andMe and the Genetic Epidemiology Research on Adult Health and Aging (GERA) cohorts, with subsequent replication and meta-analysis with data previously reported from the Consortium for Refractive Error and Myopia (CREAM).

Results

Association Results.

Analyses were restricted to subjects of European ancestry (Extended Data Figure 1) and combined results from quantitative measures of spherical equivalent and categorical myopia status. Spherical equivalent quantifies refractive error; a negative spherical equivalent, below a certain threshold defines myopia. We used results obtained from GWAS of directly measured spherical equivalent in 102,117 population-based UK Biobank participants[8], and 34,998 subjects participating in the GERA Study[9] and combined them with results of analyses of self-reported myopia in 106,086 cases and 85,757 controls from the customer base of 23andMe, Inc. (Mountain View, CA), a personal genomics company[10]. Additionally, we included results from an analysis on the refractive status inferred using demographic and self-reported information on age at first use of prescription glasses among the UK Biobank participants not contributing to the quantitative GWAS (108,956 likely myopes to 70,941 likely non-myopes, see Online Methods). All analyses were adjusted for age, sex and main principal components.

To obtain an overall association with refractive error, we meta-analyzed the results from all studies by using the z-scores from the GWAS of the spherical equivalent and the negative values of z-scores from the case-control studies (23andMe and UK Biobank), since myopia is negatively correlated with spherical equivalent. As expected, the large total sample size of the discovery meta-analysis (N=508,855) led to a nominally large genomic inflation factor (λ=1.94). The LD score regression intercept was (1.17), and the (intercept-1)/(mean(chi^2)-1) ratio of 0.097 is fully in line with the expectations of polygenicity[11].

We found associations for 438 discrete genomic regions (Figure 1, Supplementary Table 1), defined by markers contiguously associated at conventional level of GWAS significance[12,13] of $p<5x10^{-08}$, separated by more than 1 Mbp from other GWAS-associated markers, as recommended elsewhere[14]. Among them, 308 loci, including 14 on chromosome X, were not described in previous GWAS studies of refractive error[7]. The observed effect sizes were consistent across all the studies (Supplementary Table 1 and Supplementary Data 1). The association with refractive error was statistically strongest for rs12193446 ($p=9.87x10^{-328}$), within *LAMA2*, a gene previously associated with refractive error[5,6], mutations of which cause muscular dystrophy[15]. Consistent with these *LAMA2* properties, polymorphisms located within the genes coding for both major *LAMA2* receptors, *DAG1*[16] ($p=1.67x10^{-08}$ for rs111327216) and *ITGA7*[17] ($p=8.57x10^{-09}$ for rs17117860) which are also known causes of muscular dystrophy[18,19], were significantly associated with refractive error in the discovery meta-analysis.

We compared our discovery meta-analysis findings with GWAS results from 34,079 participants in the CREAM consortium, who were part of a previously reported meta-analysis[7]. To avoid any potential overlap with the UK Biobank participants, only non-UK European CREAM participants were used for replication. Despite the vast power differential, 55 of the SNPs that showed the strongest association in their respective regions in the discovery meta-analysis were significant after Bonferroni correction in the replication sample. A further 142 had a false discovery rate (FDR) < 0.05 and 192 were nominally significant at P < 0.05 (Supplementary Table 2). The effect sizes observed in the discovery and replication samples were strongly correlated (Pearson's r=0.91, Extended Data Figure 2). Meta-analysis of all five cohorts (discovery and replication) expanded the number to 449 associated of regions of variable length and number of SNPs (Extended Data Figure 3), of which 336 regions were novel (Supplementary Table 3).

Most of the 449 refractive error-associated regions contained at least one gene linked to severe ocular manifestations in the Online Mendelian Inheritance In Man (OMIM) resource or other genes with interesting link to eye disease (Supplementary Table 4). Although most loci identified through our meta-analyses were novel, several of them hosted genes that harbor mutations leading to myopia or other refractive error phenotypes (Supplementary Data 2). Several genes significantly associated with refractive error were linked to Mendelian disorders affecting corneal structure, some of which code for transcription factors involved in corneal development[20] (Supplementary Table 5). Mutations in these genes cause corneal dystrophies (*SLC4A11*, $p=5.81x10^{-11}$ for rs41281858, *TCF4*, $p=4.14x10^{-08}$, rs41396445; *LCAT*, $p=1.26x10^{-10}$, rs5923; and *DCN*, $p=3.67x10^{-09}$, rs1280632), megalocornea (*LTBP2*, $p=1.91x10^{-24}$, rs73296215) and keratoconus (*FNDC3B*, $p=1.89x10^{-14}$, rs199771582, previously described[7]). Eleven refractive error-associated genes were linked to anomalies of the crystalline lens (Supplementary Table 6), including genes linked to autosomal dominant cataracts (*PAX6* previously linked to myopia[21], $p=8.31x10^{-11}$, rs1540320; *PITX3*, $p=1.05x10^{-10}$, rs7923183; *MAF*, $p=5.50x10^{-09}$, rs16951312; *CHMP4B*, $p=9.95x10^{-11}$, rs6087538; *TDRD7*, $p=4.79x10^{-08}$, rs13301794) and lens ectopia (*FBN1*, $p=3.30x10^{-24}$, rs2017765; *ADAMTSL4*, $p=8.19x10^{-14}$, rs12131376). Some of the genes affected several eye components. For example, *LTBP2 variants* are also associated with congenital glaucoma[22], and *COL4A3* (rs7569375, $p=1.14x10^{-08}$) causes Alport syndrome, which manifests with abnormal lens shape (lenticonus) and structural changes in the retina.

Association was also observed within or near 13 genes known to harbor mutations causing microphthalmia (Supplementary Table 7), including *TENM3* ($p=2.48x10^{-11}$, rs35446926); *OTX2* ($p=6.15x10^{-11}$, rs928109); *VSX2*, ($p=4.60x10^{-10}$, rs35797567); *MFRP*, ($p=2.85x10^{-16}$, rs10892353) and the previously identified[6] *TMEM98*, ($p=3.49x10^{-43}$, rs62067167). Association was also found for *VSX1* ($p=4.59x10^{-08}$ for rs6050351), a gene that is closely regulated by VSX2[23] and believed to play important roles in eye development[24]. Many of the genes nearest associated SNPs have been linked to inherited retinal disease (Supplementary Table 8), including 32 genes linked to cone-rod dystrophies, night blindness and retinitis pigmentosa, and age-related macular degeneration (*HTRA1/ARMS2*). Among genes in novel regions associated with refractive error, *ABCA4* ($p=3.20x10^{-10}$ for rs11165052), and ARMS2/HTRA1 ($p=5.72x10^{-23}$ for rs2142308) are linked to macular disorders and numerous others to retinitis pigmentosa, retinal dystrophy and other retinal diseases, such as *FBN2*, ($p=8.63x10^{-11}$, rs6860901) , *TRAF3IP1* ($p=5.71x10^{-16}$, rs7596847), CWC27 ($p=1.84x10^{-18}$, rs1309551). Significant association was found near other genes of interest such as *DRD1* ($p=4.51x10^{-16}$, rs13190379), a dopamine receptor. Together, these results are consistent with previous suggestions of light transmission and transduction in refractive error[7,25].

Wnt signaling has previously been implicated in experimental myopia[26]. We found significant association near several *Wnt* protein-coding genes (*WNT7B*, a gene previously associated with axial length[27], $p=1.42x10^{-26}$ for rs73175083; *WNT10A*, previously associated with central corneal thickness[28], $p=1.65x10^{-17}$ for rs121908120 and *WNT3B*, $p=8.52x10^{-16}$ for rs70600), suggesting that organogenesis through *Wnt* signaling is likely to be involved in refractive error. Significant association were found at genes coding for key canonical (e.g. rs13072632 within the *CTNNB1* gene, $p=7.30x10^{-27}$; *AXIN2*, rs9895291, $p=1.40x10^{-08}$) and non-canonical *Wnt* pathway members (*NFATC3*, rs147561310, $p=1.493x10^{-12}$) or at genes coding for both (*RHOA*, rs7623687, $p=1.81x10^{-11}$ or the previously described[7] *TCF7L2*, rs56299331, $p=9.38x10^{-46}$; Supplementary Table 9).

Similar to previous published analyses[25], we found associations for genes involved in sodium, potassium, calcium magnesium and other cation transporters (Supplementary Table 10). The involvement of genes related to glutamatergic synaptic transmission was also notable (Supplementary Table 11). Glutamate is a first synapse transmitter released by photoreceptors towards bipolar cells and is the main excitatory neurotransmitter of the retina, and expression of genes participating in glutamate signaling pathways is significantly altered in myopia models[29]. These associations support the involvement in refractive error pathogenesis of neurotransmission and neuronal depolarization and hyperpolarization that was also suggested before[7]. Associations with *POU6F2* gene intronic variants (rs2696187, $p=1.11x10^{-11}$) also suggests involvement of factors related to development of amacrine and ganglion cells[30]. Other genes at refractive error-associated loci were annotated to infantile epilepsy, microcephaly, severe learning difficulty, or other inborn diseases affecting the central nervous system (CNS) in OMIM (Supplementary Table 12).

Polymorphisms in genes linked to oculocutaneous albinism (OCA) were significantly associated with refractive error (Supplementary Table 13), although typically association was found for SNPs not strongly associated with other pigmentation traits[31]. Strong association with refractive error was found near the *OCA2* gene causing OCA type 2 ($p=1.37x10^{-15}$, rs79406658), OCA3 (*TYRP1*, $p=1.18x10^{-11}$, rs62538956), OCA5 (*SLC39A8*, $p=4.03x10^{-17}$, rs13107325), OCA6 (*C10orf11*, $p=1.73x10^{-16}$, rs12256171). In addition, significant association was found near genes linked to ocular albinism (OA) on chromosome X (*TBL1X* and *GPR143*[32], $p=2.20x10^{-18}$, rs34437079) and Hermansky-Pudlak Syndrome albinism (*BLOC1S1,* $p=2.4610^{-22}$, for rs80340147; note that this gene forms a conjoint read-through transcript the *BLOC1S1-RDH5* with *RDH5*). Other associated markers were located within genes involved in systemic

pigmentation also previously associated with refractive error[7], such as *RALY* ($p=3.14x10^{-18}$, rs2284388), *TSPAN10* ($p=2.22x^{-50}$, rs9747347), as well as melanoma (*MCHR2,* $p=2.37x10^{-15}$ for rs4839756).

Functional properties of the associated markers

Among the significantly associated markers, 367 unique markers were frameshift or missense variants (Supplementary Table 14). Several are non-synonymous, such as the Arg141Leu mutation (rs1048661) within *LOXL1,* a gene that causes pseudoexfoliation syndrome and glaucoma[33] and Ala69Ser (rs10490924) in *ARMS2*, associated with increased susceptibility to age-related macular degeneration[34]. Other associated variants with predicted deleterious consequences were located in several genes, such as RGR (p=6.89x10-68, rs1042454), a gene previously associated with refractive error[7,10] and also retinitis pigmentosa[35], and within the *FBN1* gene, near clusters of mutations that cause Marfan Syndrome and anterior segment dysgenesis[36].

Because the functional link between other associated variants and development of refractive error phenotypes is less obvious, we next performed gene-set enrichment analyses to identify properties that are significantly shared by genes identified by the meta-analysis. An enrichment analysis of Gene Ontology processes (Supplementary Table 15) found enrichment for genes participating in RNA Polymerase II transcription regulation ($p=1x10^{-06}$) and nucleic acid binding transcription factor activity ($p=1.10x10^{-06}$), suggesting that many of the genetic associations we identified interfere with gene expression. “Eye development” ($p=6.10x10^{-06}$) and “Circadian regulation of gene expression” ($p=1.10x10^{-04}$) were also significantly enriched.

A transcription factor binding site (TFBS) enrichment analysis identified significant (FDR < 0.05) over-representation of sites targeted by *GATA4*, *EP300*, *RREB1,* for which association was observed in the meta-analyses (Supplementary Table 16). Binding sites of transcription factors involved in eye morphogenesis and development such as *MAF* (whose mutations cause autosomal cataract), *FOXC1* and *PITX2* (anterior segment dysgenesis) or *CRX* (cone-rod dystrophy) were also enriched. *CRX* and *PAX4*, binding sites were also significantly enriched; these transcription factors are two of the regulators of circadian rhythm and melatonin synthesis[37] alongside *OTX2*, for which SNP significant association was observed in our refractive error meta-analysis ($p=6.15x10^{-11}$ for rs928109). All of these enriched gene-sets are observed for the first time in a GWAS analysis, although the presence of some of the mechanisms that relate them to refractive error and myopia were hypothesized before[38].

Many of the variants associated with refractive error in our analyses were located within or near genes that are expressed in numerous body tissues (Extended Data Figure 4), and in particular from the nervous system, consistent with our evidence of extraocular, central nervous system involvement in refractive error. Within the eye, these genes were particularly strongly expressed in eye tissues such as cornea, ciliary body, trabecular meshwork[39] and retina[40] (Extended Data Figure 5, Supplementary Table 17). A stratified LD score regression applied to specifically expressed genes (LDSC-SEG)[41] revealed the results of the GWAS are most strongly correlated with genes expressed in the retina and basal ganglia in the central nervous system but these correlations are not significant after multiple testing correction (Extended Data Figure 6 and Supplementary Table 18). It is possible that the strength of these correlations was constrained by the fact that in most cases, available expression levels were measured in adult samples, while refractive error and myopia are primarily developed in younger ages.

A Summary data-based Mendelian Randomization (SMR) analysis[42] integrating GWAS with eQTL data from peripheral blood[43] and brain tissues[44] found concomitant association with refractive error and

eQTL transcriptional regulation effects for 159 and 97 genes respectively (Supplementary Tables 19 and 20). A similar analysis integrating GWAS summary data with methylation data from brain tissues found association with both refractive error and changes in methylation for 134 genes (Supplementary Table 21).

Genetic effects shared between refractive error and other conditions

Examining the GWAS Catalog[45], some of the genetic variants reported here were previously associated with refractive error, and with other traits, in particular intraocular pressure, intelligence and education; the latter two are known myopia risk factors (Supplementary Table 22). We used LD score regression to assess the correlation of genetic effects between refractive error and other phenotypes from GWAS summary statistics (Supplementary Table 23). refractive error genetic risk was significantly correlated with intelligence, both in childhood[46] (r_g=-0.27, p=4.76x10^{-09}) and adulthood (fluid intelligence score r_g=-0.25, p=1.56x10^{-39}), educational attainment (defined as the number of years spent in formal education, r_g=-0.24, p=3.36x10^{-54}), self-reported cataract (r_g=-0.31, p=4.70x10^{-10}) and intraocular pressure (IOP, r_g=-0.14, p=1.04x10^{-12}).

Higher educational attainment appears to cause myopia as demonstrated by Mendelian randomization (MR) studies[47]. A gene by environment interaction GWAS for spherical equivalent and educational attainment (using age at completion of formal full-time education as a proxy) was conducted in 66,242 UK Biobank participants. Despite the relatively well-powered sample, only one locus yielded evidence of statistically significant interaction (rs536015141 within *TRPM1*, p=2.35x10^{-09}, Supplementary Table 24), suggesting that the true relationship between refractive error and education is compounded by several factors and may not be linear in nature, as suggested recently[48]. TRPM1 is localized in rod ON bipolar cell dendrites, and rare mutations cause congenital stationary night blindness[49], often associated with high myopia.

To further explore the nature of the relationship between refractive error and IOP, we built MR models using genetic effects previously reported for IOP[50]. On average, every 1 mmHg increase in IOP predicts a 0.05-0.09 diopters decrease in spherical equivalent (Supplementary Table 25, Extended Data Figure 7). We also built a MR model to assess the relationship between intelligence and spherical equivalent, but statistical evidence in this case points towards genetic pleiotropy rather than causation (Supplementary Table 26). This suggests that both myopia and intelligence are often influenced by the same factors, but without direct causal path linking one to the other. We found no significant genetic correlations between refractive error and the glaucoma endophenotype vertical cup to disc ratio (r_g=-0.01, p=0.45), or hair pigmentation (r_g=-0.03, p=0.35). Therefore, refractive error and pigmentation may have different allelic profiles with limited sharing of genetic risk.

Conditional analysis and risk prediction

We subsequently carried out a conditional analysis[51] on the meta-analysis summary results and found a total of 904 independent SNPs significantly associated with refractive error. 890 of these markers were available in the EPIC-Norfolk Study, an independent cohort that did not participate in the refractive error meta-analysis (Extended Data Figure 8). These markers alone explained 12.1% of the overall spherical equivalent phenotypic variance in a regression model or 18.4% (SE=0.04) of the spherical equivalent heritability. Newly associated markers found in our meta-analysis, but not in the previous large GWAS[7],

explain 4.6% (SE=0.01) of the spherical equivalent phenotypic variance in EPIC-Norfolk Study, which is an improvement of one third compared to heritability explained by previously associated markers[7].

Predictive models, based on the above-mentioned 890 SNPs, along with age and sex, were predictive of myopia (versus all non-myopia controls) with areas under the receiving operating characteristic curve (AUC) of 0.67, 0.74 and 0.75 (Figure 2), depending on the severity cutoff for myopia ($\leq$ -0.75D, $\leq$ -3.00D and $\leq$ -5.00D respectively). The performance of the predictions appears not to improve for myopia definitions of -3.00D or worse, suggesting that the information extracted from our meta-analysis is more representative of the genetic risk for common myopia seen in the general population, than for more severe forms of myopia, which may have a distinctive genetic architecture.

Further exploration of refractive error genetic architecture

Using information from over half a million population-based participants SNPs identified in these analyses still only explain 18.4% of the spherical equivalent heritability. We next assessed how many common SNPs are likely to explain the entire heritable component of refractive error, and what sample sizes are likely to be needed in the future to identify them, using the likelihood-based approach described elswhere[52]. We estimate that approximately 13,808 (SE=969) polymorphic variants are likely to be behind the full refractive error heritability. Similar to other quantitative phenotypic traits that are previously published[52], our analyses estimate that 10.3% (SE=1.0%) of the phenotypic variance is likely explained by a batch of approximately 543 (SE=81) common genetic variants of relatively large effect size and a further 20.8% (SE=0.9%) of the entire phenotypic variance explained by the remainder. With increased sample sizes, we project that the proportion of variance explained will continue to improve fast but will start plateauing for sample sizes above one million, after which further increases in sample size will likely yield ever diminishing additional phenotypic variance (Extended Data Figure 9).

Discussion

Our results provide evidence for at least two major sets of mechanisms in the pathogenesis of refractive error. The first affect intraocular pressure, eye structure, ocular development and physiology, and the second are CNS-related, including circadian rhythm control. Contributors to refractive error include all anatomical factors that alter refractive power relative to eye size, light transmittance, photoconductance and higher cerebral functions.

The findings implicate almost every single anatomical components of the eye, which along with the central nervous system participate in the development of refractive error. The healthy cornea contributes to 70% of the optical refractive power of the eyes[53] and genes involved in corneal structure, topography and function may directly contribute to refractive error through direct changes in the corneal refraction. Our results show that several genes involved in lens development also contribute to refractive error in the general population. It is unclear if their contribution is mediated through alterations in biomechanical properties that affect eyes' ability to accommodate, changes to the lens refractive index, or alterations in light transmission properties that impair the ability to focus images on the retina.

Many retinal genes are implicated in the development of refractive error, reflecting the role of light in mediating eye growth and the importance of the retina's role in light transduction and processing[7]. Associations with refractive error at genes coding for gated ion channels and glutamate receptors point to the photoreceptor-bipolar cell interface as a potentially key factor in refractive error. Rare mutations

in several of our associated genes cause night blindness, implicating the rod system in the pathophysiology of refractive error, but many also affect cone pathways. The *TRPM1* gene, important for rod ON bipolar cell polarity[54], is also implicated in the gene-education interaction analysis. Associations observed for the *VSX1* and VSX2, its negative regulator, genes implicate the cone bipolar cells[55].

The association with genes involved in pigmentation, including most of the OCA-causing genes, raises questions about the relationship between melanin, pigmentation and eye growth and development. These associations are unlikely to be influenced by any cryptic population structure in our samples, which our analyses were designed to control. None of the major pigmentation-associated SNPs[31] was directly associated with refractive error and there was no significant correlation of genetic effects between refractive error and pigmentation.

The mechanisms linking pigmentation with refractive error are unclear. Foveal hypoplasia[56] and optic disc[57] dysplasias are common in all forms of albinism[58]. Although melanin synthesis is disrupted in albinism, both melanin and dopamine are synthesized through shared metabolic pathways. Disc and chiasmal lesions in albinism are often attributed to dopamine[59], but we found limited evidence supporting an association with refractive error for genetic variants involved in dopamine signaling. The scarcity of association with refractive error for genes involved in dopamine-only pathways contrasts with the abundance of association for genes involved in pigmentation and melanin synthesis. This may suggest that melanin metabolism is connected to refractive error through other mechanisms that are independent from the metabolic pathways it shares with dopamine production. Melanin reaches the highest concentrations in the retinal pigment epithelium at the outmost layer of the retina, and anteriorly, in the iris and variations in pigmentation may affect the intensity of the light reaching the retina. Light exposure is a major protective factor for development of myopia[60,61] It is possible that pigmentation plays a role in light signal transmission and transduction.

Animal model experiments suggest that in addition to local ocular mechanisms, emmetropization (the process by which the eye develops to minimize refractive error) is strongly influenced by the CNS[62]. The strong correlation of genetic risks between refractive error and intelligence and association found for genes linked to severe learning disability support an involvement of the CNS in emmetropization and refractive error pathogenesis.

Results from gene-set enrichment analysis demonstrate an interesting evolution with increasing sample sizes. While smaller previous studies were sufficiently powered to discover enrichment of low, cell-level properties, such as cation channel activity and participation in the synaptic space structures[25], significantly more powered recent studies have found additional evidence for enrichment and involvement of more integrated physiological functions, such as light signal processing in retinal cells and others[7]. Beyond the identification of a much larger number of genes and explaining significantly higher proportions of heritability, our results, based in a considerably more statistically powered sample, uphold the previous findings and support the involvement of the same molecular and physiological mechanisms that were previously described.

In line with expectations from a higher power of association to discover genes and gene sets individually responsible for even smaller proportions of the refractive error variance[63], we find evidence for even higher regulatory mechanisms, that act more holistically over the eye development or integrate eye growth and homeostasis with other processes of extraocular nature. For example, we found evidence that binding sites of transcription factors involved in the control of circadian rhythm are significantly enriched among genes associated with refractive error. Circadian rhythm is important in emmetropization and its disruption leads to myopia in animal knock-out models[38], potentially through dopamine-mediated mechanisms, or changes in IOP and diurnal variations.

Most of the loci identified through our meta-analysis are not subject to particularly strong and systematic evolutionary pressures (Extended Data Figure 10). The variability in minor allele frequencies observed across loci associated with refractive error may therefore be the result of genetic drift. However, given the variety of the different visual components whose disruptions can result in refractive error, this variability may also be the result of overall balancing forces which encourage high allelic diversity of genes involved in refractive error, providing additional buffering capacity to absorb environmental pressures[48] or genetic disruptions on any of the individual components of the visual system.

Our results cast light on potential mechanisms that contribute to refractive error in the general population and have identified the genetic factors that explain a considerable proportion of the heritability and phenotypic variability of refractive error. This allows us to improve significantly our ability to make predictions of myopia risk and generate novel hypotheses on how multiple aspects of visual processing affect emmetropization, which may pave the way to personalized risk management and treatment of refractive error in the population in the future.

Acknowledgements

P.T.K. and P.J.F oversaw the UK Biobank eye data acquisition with support from The National Institute for Health Research (NIHR); Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology. The UK Biobank Eye and Vision Consortium was supported by grants from UK NIHR (BRC3_026), Moorfields Eye Charity (ST 15 11 E), Fight for Sight (1507/1508), The Macular Society, The International Glaucoma Association (IGA, Ashford UK) and Alcon Research Institute. V.V. is supported by a core UK Medical Research Council grant MC_UU_00007/10.

23andMe thanks research participants and employees of 23andMe for making this work possible (list of contributing staff in the Supplementary Note).

Genotyping of the GERA cohort was funded by the US National Institute on Aging; National Institute of Mental Health and National Institute of Health Common Fund (RC2 AG036607); data analyses by the National Eye Institute (NEI R01 EY027004, E.J.), National Institute of Diabetes and Digestive and Kidney Diseases (R01 DK116738, E.J.).

The CREAM GWAS meta-analysis was supported by European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Programme (grant 648268 to C.C.W.K), the Netherlands Organisation for Scientific Research (NWO, 91815655 to C.C.W.K) and the National Eye Institute (R01EY020483). V.J.M.V. acknowledges funding from the Netherlands Organisation for Scientific Research (NWO, grant 91617076).

S.M. acknowledges support from the National Health and Medical Research Council (NHMRC) of Australia (grants 1150144, 1116360, 1154543, 1121979).

EPIC-Norfolk infrastructure and core functions are supported by the Medical Research Council (G1000143) and Cancer Research UK (C864/A14136). Genotyping funded by the UK MRC (MC_PC_13048). A.K.P. is supported by a Moorfields Eye Charity grant. P.J.F. received support from the Richard Desmond Charitable Trust, the National Institute for Health Research to Moorfields Eye Hospital and the Biomedical Research Centre for Ophthalmology.

RW and PGH were supported by the National Eye Institute of the National institutes of Health under award number R21EY029309. M.J.S. is a recipient of a Fight for Sight PhD studentship. K.P. is a recipient of a Fight for Sight PhD studentship. P.G.H. the recipient of a FfS ECI fellowship. P.G.H. and C.J.H. acknowledge the TFC Frost Charitable Trust Support for the KCL Department of Ophthalmology. Statistical analyses were run in King's College London Rosalind HPC LINUX Clusters and cloud servers. The UK Biobank data was accessed as part of the UK Biobank projects 669 and 17615.

J.S.R. is supported in part by the NIHR Biomedical Research Centres at Moorfields Eye Hospital/UCL Institute of Ophthalmology and at the UCL Institute of Child Health/Great Ormond Street Hospital and is an NIHR Senior Investigator. P.M.C. was funded by the Ulverscroft Foundation. O.A.M is supported by Wellcome Trust grant 206619_Z_17_Z and the NIHR Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology.

Author Contributions

P.G.H., J.S.R., E.J. and C.J.H. conceived and designed the study. P.T.K., P.J.F., J.S.R. contribute to the collection of data. P.G.H., H.C., A.P.K., R.W., M.S.T., J.Y., K.K.T., P.M.C., V.V., J.A.G and E.J. performed statistical analysis. A.P.K., M.J.S., K.P., K.K.T., AS and J.A.G. performed post-GWAS follow-up analyses. P.G.H., H.C., A.P.K., R.W., J.S.R., E.J., C.J.H. wrote the manuscript with help from O.A.M., P.M.C., R.B.M., V.J.M.V., A.S., R.A.S., N.W., A.W.H., D.A.M., C.C.W.K., S.M., P.T.K., P.J.F. and J.A.G. who helped with the interpretation of the results.

Competing interests

23andMe is a consumer genomics company.

References:

1. Vos, T. *et al.* Global, regional, and national incidence, prevalence, and years lived with disability for 328 diseases and injuries for 195 countries, 1990–2016: a systematic analysis for the Global Burden of Disease Study 2016. *The Lancet* **390**, 1211-1259 (2017).
2. WHO. The Global Burden of Disease. 2004 Update ISBN-13: 9789241563710. *ISBN-10* **651629118**(2008).
3. Williams, K.M. *et al.* Increasing Prevalence of Myopia in Europe and the Impact of Education. *Ophthalmology* **122**, 1489-97 (2015).
4. Sanfilippo, P.G., Hewitt, A.W., Hammond, C.J. & Mackey, D.A. The heritability of ocular traits. *Surv Ophthalmol* **55**, 561-83 (2010).
5. Kiefer, A.K. *et al.* Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. *PLoS Genet* **9**, e1003299 (2013).
6. Verhoeven, V.J. *et al.* Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet* **45**, 314-8 (2013).
7. Tedja, M.S. *et al.* Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. *Nat Genet* **50**, 834-848 (2018).

8. Cumberland, P.M. *et al.* Frequency and Distribution of Refractive Error in Adult Life: Methodology and Findings of the UK Biobank Study. *PLoS One* **10**, e0139780 (2015).
9. Kvale, M.N. *et al.* Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1051-60 (2015).
10. Pickrell, J.K. *et al.* Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* **48**, 709-17 (2016).
11. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
12. Dudbridge, F. & Gusnanto, A. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol* **32**, 227-34 (2008).
13. Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M.J. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* **32**, 381-5 (2008).
14. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).
15. Oliveira, J. *et al.* LAMA2 gene analysis in a cohort of 26 congenital muscular dystrophy patients. *Clin Genet* **74**, 502-12 (2008).
16. Colognato, H. *et al.* Identification of dystroglycan as a second laminin receptor in oligodendrocytes, with a role in myelination. *Development* **134**, 1723-36 (2007).
17. Burkin, D.J. & Kaufman, S.J. The alpha7beta1 integrin in muscle development and disease. *Cell Tissue Res* **296**, 183-90 (1999).
18. Ervasti, J.M. & Campbell, K.P. Dystrophin-associated glycoproteins: their possible roles in the pathogenesis of Duchenne muscular dystrophy. *Mol Cell Biol Hum Dis Ser* **3**, 139-66 (1993).
19. Mayer, U. *et al.* Absence of integrin alpha 7 causes a novel form of muscular dystrophy. *Nat Genet* **17**, 318-23 (1997).
20. Jean, D., Ewan, K. & Gruss, P. Molecular regulators involved in vertebrate eye development. *Mech Dev* **76**, 3-18 (1998).
21. Hammond, C.J., Andrew, T., Mak, Y.T. & Spector, T.D. A susceptibility locus for myopia in the normal population is linked to the PAX6 gene region on chromosome 11: a genomewide scan of dizygotic twins. *Am J Hum Genet* **75**, 294-304 (2004).
22. Ali, M. *et al.* Null mutations in LTBP2 cause primary congenital glaucoma. *Am J Hum Genet* **84**, 664-71 (2009).
23. Clark, A.M. *et al.* Negative regulation of Vsx1 by its paralog Chx10/Vsx2 is conserved in the vertebrate retina. *Brain Res* **1192**, 99-113 (2008).
24. Heon, E. *et al.* VSX1: a gene for posterior polymorphous dystrophy and keratoconus. *Hum Mol Genet* **11**, 1029-36 (2002).
25. Hysi, P.G. *et al.* Common mechanisms underlying refractive error identified in functional analysis of gene lists from genome-wide association study results in 2 European British cohorts. *JAMA Ophthalmol* **132**, 50-6 (2014).
26. Ma, M. *et al.* Wnt signaling in form deprivation myopia of the mice retina. *PLoS One* **9**, e91086 (2014).
27. Miyake, M. *et al.* Identification of myopia-associated WNT7B polymorphisms provides insights into the mechanism underlying the development of myopia. *Nat Commun* **6**, 6689 (2015).
28. Cuellar-Partida, G. *et al.* WNT10A exonic variant increases the risk of keratoconus by decreasing corneal thickness. *Hum Mol Genet* **24**, 5060-8 (2015).
29. Stone, R.A. *et al.* Image defocus and altered retinal gene expression in chick: clues to the pathogenesis of ametropia. *Invest Ophthalmol Vis Sci* **52**, 5765-77 (2011).

30. Zhou, H., Yoshioka, T. & Nathans, J. Retina-derived POU-domain factor-1: a complex POU-domain gene implicated in the development of retinal ganglion and amacrine cells. *J Neurosci* **16**, 2261-74 (1996).
31. Hysi, P.G. *et al.* Genome-wide association meta-analysis of individuals of European ancestry identifies new loci explaining a substantial fraction of hair color variation and heritability. *Nat Genet* **50**, 652-656 (2018).
32. Fabian-Jessing, B.K. *et al.* Ocular albinism with infertility and late-onset sensorineural hearing loss. *Am J Med Genet A* **176**, 1587-1593 (2018).
33. Thorleifsson, G. *et al.* Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. *Science* **317**, 1397-400 (2007).
34. Rivera, A. *et al.* Hypothetical LOC387715 is a second major susceptibility gene for age-related macular degeneration, contributing independently of complement factor H to disease risk. *Hum Mol Genet* **14**, 3227-36 (2005).
35. Morimura, H., Saindelle-Ribeaudeau, F., Berson, E.L. & Dryja, T.P. Mutations in RGR, encoding a light-sensitive opsin homologue, in patients with retinitis pigmentosa. *Nat Genet* **23**, 393-4 (1999).
36. Robinson, P.N. *et al.* Mutations of FBN1 and genotype-phenotype correlations in Marfan syndrome and related fibrillinopathies. *Hum Mutat* **20**, 153-61 (2002).
37. Rohde, K., Moller, M. & Rath, M.F. Homeobox genes and melatonin synthesis: regulatory roles of the cone-rod homeobox transcription factor in the rodent pineal gland. *Biomed Res Int* **2014**, 946075 (2014).
38. Chakraborty, R. *et al.* Circadian rhythms, refractive development, and myopia. *Ophthalmic Physiol Opt* **38**, 217-245 (2018).
39. Carnes, M.U., Allingham, R.R., Ashley-Koch, A. & Hauser, M.A. Transcriptome analysis of adult and fetal trabecular meshwork, cornea, and ciliary body tissues by RNA sequencing. *Exp Eye Res* **167**, 91-99 (2018).
40. Ratnapriya, R. *et al.* Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat Genet* **51**, 606-610 (2019).
41. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).
42. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
43. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-1243 (2013).
44. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun* **9**, 2282 (2018).
45. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
46. Benyamin, B. *et al.* Childhood intelligence is heritable, highly polygenic and associated with FNBP1L. *Mol Psychiatry* **19**, 253-8 (2014).
47. Mountjoy, E. *et al.* Education and myopia: assessing the direction of causality by mendelian randomisation. *BMJ* **361**, k2022 (2018).
48. Pozarickij, A. *et al.* Quantile regression analysis reveals widespread evidence for gene-environment or gene-gene interactions in myopia development. *Commun Biol* **2**, 167 (2019).
49. Audo, I. *et al.* TRPM1 is mutated in patients with autosomal-recessive complete congenital stationary night blindness. *Am J Hum Genet* **85**, 720-9 (2009).

50. Khawaja, A.P. *et al.* Genome-wide analyses identify 68 new loci associated with intraocular pressure and improve risk prediction for primary open-angle glaucoma. *Nat Genet* **50**, 778-782 (2018).
51. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).
52. Zhang, Y., Qi, G., Park, J.H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat Genet* **50**, 1318-1326 (2018).
53. Zadnik, K. *et al.* Normal eye growth in emmetropic schoolchildren. *Optom Vis Sci* **81**, 819-28 (2004).
54. Li, Z. *et al.* Recessive mutations of the gene TRPM1 abrogate ON bipolar cell function and cause complete congenital stationary night blindness in humans. *Am J Hum Genet* **85**, 711-9 (2009).
55. Chow, R.L. *et al.* Vsx1, a rapidly evolving paired-like homeobox gene expressed in cone bipolar cells. *Mech Dev* **109**, 315-22 (2001).
56. Struck, M.C. Albinism: Update on ocular features. *Current Ophthalmology Reports* **3**, 232-237 (2015).
57. Mohammad, S. *et al.* Characterization of Abnormal Optic Nerve Head Morphology in Albinism Using Optical Coherence Tomography. *Invest Ophthalmol Vis Sci* **56**, 4611-8 (2015).
58. Yahalom, C. *et al.* Refractive profile in oculocutaneous albinism and its correlation with final visual outcome. *Br J Ophthalmol* **96**, 537-9 (2012).
59. Lopez, V.M., Decatur, C.L., Stamer, W.D., Lynch, R.M. & McKay, B.S. L-DOPA is an endogenous ligand for OA1. *PLoS Biol* **6**, e236 (2008).
60. Karouta, C. & Ashby, R.S. Correlation between light levels and the development of deprivation myopia. *Invest Ophthalmol Vis Sci* **56**, 299-309 (2014).
61. Wu, P.C., Tsai, C.L., Wu, H.L., Yang, Y.H. & Kuo, H.K. Outdoor activity during class recess reduces myopia onset and progression in school children. *Ophthalmology* **120**, 1080-5 (2013).
62. Troilo, D., Gottlieb, M.D. & Wallman, J. Visual deprivation causes myopia in chicks with optic nerve section. *Curr Eye Res* **6**, 993-9 (1987).
63. de Leeuw, C.A., Neale, B.M., Heskes, T. & Posthuma, D. The statistical properties of gene-set analysis. *Nat Rev Genet* **17**, 353-64 (2016).

Figure Legends

Figure 1. All GWAS-associated regions from the main meta-analysis. Each band is a true scale of genomic regions associated with refractive error listed in Supplementary Table 1 (+250kbp on each side to make smaller regions more visible). The different color codes represent the significance (p-value) for the genetic variant within that region that displays the strongest evidence for association.

Figure 2. Receiver Operating Characteristic (ROC) curves for myopia predictions, using information from 890 SNP markers identified in the meta-analysis. The three different colors represent three different curves for each of the different definition of myopia: green – all myopia (< -0.75D), magenta – moderate myopia (< -3.00 D) and brown - severe myopia (defined as < -5.00 D).

Online Methods

Study Participants

The UK Biobank

The UK Biobank is a multisite cohort study of UK residents aged 40 to 69 years who were registered with the National Health Service (NHS) and living up to 25 miles from a study center. Detailed study protocols are available online (http://www.ukbiobank.ac.uk/resources/ and http://biobank.ctsu.ox.ac.uk/crystal/docs.cgi). It was conducted with the approval of the North-West Research Ethics Committee (ref 06/MRE08/65), in accordance with the principles of the Declaration of Helsinki, and all participants gave written informed consent.

Two separate groups of UK Biobank participants were included in these analyses. The first included participants whose refractive error was directly measured (non-cycloplegic autorefraction using the Tomey RC 5000 Auto Refkeratometer, Tomey Corp., Nagoya, Japan). Direct measurements of refractive errors were available for 22.7% of the UK Biobank sample. To ensure reliable and accurate refractive error data, previously published QC criteria were applied[64]. The spherical equivalent was calculated as spherical refractive error (UK Biobank codes 5084 and 5085) plus half the cylindrical error (UK Biobank 5086 and 5087) for each eye.

The second UK Biobank group included participants without direct measurement of refractive error. These participants refractive error status was inferred using questionnaire and other indirect data. Available demographic and clinical information were used to obtain an estimate about the individual's likely myopia status. A Support Vector Machine (SVM) model, with age, sex, age of first spectacle wear and year of birth as prediction parameters was used to infer participants' myopia status. Initial training took place in 80% randomly selected UK Biobank participants of European descent for whom direct spherical equivalent and refractive error status were available. Then the performance was assessed in the remaining 20% of UK Biobank participants of European descent for whom direct spherical equivalent and refractive error status were available. Finally, the SVM predictions in the remaining individuals with no direct spherical error measurements available using the model developed for the training data.

All UK Biobank genotypes were obtained as described elsewhere[65]. The UK Biobank team then performed imputation from a combined Haplotype Reference Consortium (HRC) and UK10K reference panel. Phasing on the autosomes was carried out using a modified version of the SHAPEIT2[66] program modified to allow for very large sample sizes. Only HRC-imputed variants were used for the purpose our analyses of the UK Biobank participants. The variant-level quality control exclusion metrics applied to imputed data for GWAS included the following: call rate < 95%, Hardy–Weinberg equilibrium $P < 1 \times 10^{-6}$, posterior call probability < 0.9, imputation quality < 0.4, and MAF < 0.005. The Y chromosome and mitochondrial genetic data were excluded from this analysis. In total, 10,263,360 imputed DNA sequence variants were included in our analysis. Non-European ancestry and participants with relatedness corresponding to third-degree relatives or closer, samples with excess of missing genotype calls or heterozygosity were excluded. In total, genotypes were available for 102,117 participants of European ancestry with spherical equivalent data.

Association models in the first UK Biobank subset used the average of spherical equivalent as the outcome and allele dosages at each genetic locus as predictors. Mixed linear regressions, adjusting for age, sex and the first 10 principal components, implemented in the Bolt-LMM software[67] were used.

For the second UK Biobank subset, for which no direct spherical equivalent measurement was available, the mixed linear model was built with the predicted myopia status as outcome and using the same covariates as for the previously described linear regression analysis on spherical equivalent. Odds Ratios were obtained from the beta regression coefficient using the equation:

$$\ln(\mathrm{OR}) = \frac{\beta}{\mu(1-\mu)}$$

where μ is the fraction of the cases in the sample (μ=0.606). Genotypes with MAF <0.01 and MAC< 400 were removed from analyses in this group.

23andMe

Participating subjects were all volunteers from the 23andMe (Mountain View, CA, USA) personal genomics company customer base. All participants provided informed consent and answered surveys online according to the approved 23andMe human subjects protocol, which was reviewed and approved by Ethical & Independent Review Services, a private institutional review board (http://www.eandireview.com). The participants were identified as myopia cases if they self-reported a diagnosis of myopia or suffering from symptoms of myopia (see Supplementary Notes for more detail).

DNA extraction and genotyping were performed on saliva samples by CLIA-certified and CAP-accredited clinical laboratories of Laboratory Corporation of America. Samples were genotyped on one of four genotyping platforms and batches (Illumina HumanHap550, BeadChip, SNPs, Illumina OmniExpress, plus a variable number of custom SNP assays). Only samples with more than 98.5% genotyping success rate were included. Ethnic categorization was conducted using a support vector machine (SVM) which classified individual haplotypes into one of the 31 reference populations derived from public datasets (the Human Genome Diversity Project, HapMap, and 1000 Genomes), as well as 23andMe customers who have reported having four grandparents from the same country. Genotypes were imputed against the September 2013 release of 1000 Genomes Phase1 reference haplotypes using a Beagle haplotype graph-based phasing algorithm for the autosomal and Minimac2[68] for X Chromosome loci.
Association test results were computed by linear regression assuming additive allelic effects using imputed allele dosages. Covariates for age, gender, the first ten principal components to account for residual population structure were also included into the model.

The Genetic Epidemiology Research in Adult Health and Aging (GERA) cohort

GERA is part of the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and has been described in detail elsewhere[69]. It comprises adult men and women consenting members of Kaiser Permanente Northern California (KPNC), an integrated health care delivery system, with ongoing longitudinal records from vision examinations. For this analysis, 34,998 adults (25 years and older), who self-reported as non-Hispanic white, and who had at least one assessment of spherical equivalent obtained between 2008 and 2014 were included. All study procedures were approved by the Institutional Review Board of the Kaiser Foundation Research Institute. Participants underwent vision examinations, and most subjects had multiple measures for both eyes. Spherical equivalent was assessed as the sphere + cylinder/2. The spherical equivalent was selected from the first documented

assessment, and the mean of both eyes was used. Individuals with histories of cataract surgery (in either eye), refractive surgery, keratitis, or corneal diseases were excluded from further analyses.
DNA samples from GERA individuals were extracted from Oragene kits (DNA Genotek Inc., Ottawa, ON, Canada) at KPNC and genotyped at the Genomics Core Facility of the University of California, San Francisco (UCSF). DNA samples were genotyped using the Affymetrix Axiom arrays (Affymetrix, Santa Clara, CA, USA). SNPs with initial genotyping call rate ≥97%, allele frequency difference ≤0.15 between males and females for autosomal markers, and genotype concordance rate >0.75 across duplicate samples were included. In addition, SNPs with genotype call rates <90% were removed, as well as SNPs with a minor allele frequency (MAF) < 1%.
Imputation pre-phasing of genotypes was done using Shape-IT v2.r72719[66], variants were imputed from the cosmopolitan 1000 Genomes Project reference panel (phase I integrated release; http://1000genomes.org) using IMPUTE2 v2.3.0[70]. Variants with an imputation IMPUTE $r^2 < 0.3$ were excluded, and analyses were restricted to SNPs that had a minor allele count (MAC) ≥ 20.

For each SNP locus, linear regressions of each individual's spherical equivalent were performed with the following covariates: age at first documented spherical equivalent assessment, sex, and genetic principal components using PLINK v1.9 (www.cog-genomics.org/plink/1.9/). Data from each SNP were modeled using additive dosages to account for the uncertainty of imputation. The top 10 ancestry PCs were included as covariates, as well as the percentage of Ashkenazi ancestry to adjust for genetic ancestry, as described previously[69].

The Consortium for Refractive Error And Myopia (CREAM)

All participants selected for this study were of European descent, 25 years of age or older. refractive error was represented by measurements of refraction and spherical equivalent (Spherical equivalent = spherical refractive error +1/2 cylinder refractive error) was the outcome variable for CREAM. Participants with conditions that could alter refraction, such as cataract surgery, laser refractive procedures, retinal detachment surgery, keratoconus, or ocular or systemic syndromes were excluded from the analyses. Recruitment and ascertainment strategies varied by study and were previously published elsewhere[71].

The genotyping process has been described elsewhere[71]. Samples were genotyped on different platforms, and study-specific QC measures of the genotyped variants were implemented before association analysis. Genotypes were imputed with the appropriate ancestry-matched reference panel for all cohorts from the 1000 Genomes Project (Phase I version 3, March 2012 release). Quality control criteria used for SNP and sample inclusions These metrics were similar to those described in a previous GWAS analyses and detailed information for each cohort is described elsewhere[71].

To avert sample overlap, cohorts from the United Kingdom (1985BBC, ALSPAC-Mothers, EPIC-Norfolk, ORCADES and Twins UK) were excluded from the GWAS meta-analysis. Association analyses were performed as described elsewhere[71]For each individual cohort, a single-marker analysis for the phenotype of SphE (in diopters) was carried out with linear regression with adjustment for age, sex and up to the first five principal components. For all non-family-based cohorts, one of each pair of relatives was removed. In family-based cohorts, mixed model-based tests of association were used to adjust for within-family relatedness.

The European Prospective Investigation into Cancer (EPIC) Study

The EPIC-EPIC is one of the UK arms of a broad pan-European prospective cohort study designed to investigate the etiology of major chronic diseases[72]. This study was conducted following the principles of the Declaration of Helsinki and the Research Governance Framework for Health and Social Care. The study was approved by the Norfolk Local Research Ethics Committee (05/Q0101/191) and East Norfolk & Waveney NHS Research Governance Committee (2005EC07L). All participants gave written, informed consent. Refractive error was measured in both eyes using a Humphrey Auto-Refractor 500 (Humphrey Instruments, San Leandro, California, USA). Spherical equivalent was calculated as spherical refractive error plus half the cylindrical error for each eye.
The EPIC-Norfolk participants were genotyped using the Affymetrix UK Biobank Axiom Array (the same array as used in UK Biobank); 7,117 contributed to the current study. SNP exclusion criteria included: call rate < 95%, abnormal cluster pattern on visual inspection, plate batch effect evident by significant variation in minor allele frequency, and/or Hardy-Weinberg equilibrium $P < 10^{-7}$. Sample exclusion criteria included: DishQC < 0.82 (poor fluorescence signal contrast), sex discordance, sample call rate < 97%, heterozygosity outliers (calculated separately for SNPs with minor allele frequency >1% and <1%), rare allele count outlier, and impossible identity-by-descent values. Individuals with relatedness corresponding to third-degree relatives or closer across all genotyped participants were also removed from further analyses. Following these steps all participants were of European descent. Data were pre-phased using SHAPEIT[66] version 2 and imputed to the Phase 3 build of the 1000 Genomes project[74] (October 2014) using IMPUTE[70] version 2.3.2.

The relationship between allele dosage and mean spherical equivalent was analyzed using linear regression adjusted for age, sex and the first 5 principal components. Analyses were carried out using SNPTEST version 2.5.1.

Statistical analyses

We conducted two meta-analyses. For the initial meta-analysis (discovery), we used summary statistic results from the UK Biobank 1st and 2nd subset, the GERA and 23andMe Studies.

For the final meta-analysis, we used all available information (UK Biobank 1 and 2, the GERA, 23andMe and CREAM Consortium).

For all meta-analyses we applied a Z-score method, weighted by the effective population sample size, as implemented in METAL[75]. No genomic control adjustment was applied during the meta-analysis.

The effective population size was calculated per each locus and as was equal to the total sample size if a linear regression or linear mixed model were used. For case-control studies the effective population was calculated as:

$$\mathrm{N.eff} = 2/(\frac{1}{N.c_{\text{ases}}} + \frac{1}{N.c_{\text{ontrols}}})$$

as recommended before[76], where N.eff is the effective sample size, N.cases is the number of cases considered to have myopia and N.controls is the number of subjects considered not to have myopia. Following this method, we calculated that for the full-sample analysis of 542,934 subjects, due to the presence of two case-control cohort, our effective sample sizes was 379,227.

Only SNPs with minor allele frequency of at least 1%, which were available from at least 70% of the maximum number of participants across all studies and that were not missing in more than one strata (cohorts), were considered further.

Conditional analyses were conducted using the conditional and joint analysis on summary data (COJO) as implemented in the GCTA program[77] to identify independent effects within associated loci as well as the calculation of the phenotypic variance explained[78] by all polymorphisms associated with the trait after the conditional analyses. The threshold of significance was set at $5x10^{-08}$ and the collinearity threshold was set at r^2=0.9.

Genomic inflation was assessed using the package 'gap' in R (https://cran.r-project.org/) and to distinguish between the effect of polygenicity and those arising from sample stratification or uncontrolled population admixture, the LD score regression intercepts were calculated using the program LD Score (https://github.com/bulik/ldsc).

Bivariate genetic correlations between refractive error and other complex traits whose summary statistics are publicly available were assessed following previously described methodologies[79], using the program LD Score (https://github.com/bulik/ldsc).

To assess the potential value of the loci associated with refractive error to predict myopia, regression-based models were trained and tested separately in two separate groups. The training set comprises the European UK Biobank participants for whom the spherical equivalent measurements were available. The models were tested in the EPIC-Norfolk cohort, which was not part of any of the analyses through which the genetic associations were identified.

The model in included age, sex, and the major genetic variants associated with refractive error after the conditional analysis. Three different definitions of myopia were used based on sliding spherical equivalent thresholds: $M_1 \leq$ -0.75D, $M_2 \leq$ -3.00 D and $M_3 \leq$ -5.00D. These three different definitions of myopia were chosen to correspond to the generally accepted definitions of "any myopia", "moderate myopia" and "high myopia". For the latter, we opted for the -5.00D, because definitions based on the more stringent threshold of $\leq$ -6.00D would have not allowed for a sufficient number of cases in the testing set. For the purpose of these analyses, a "control" was any subject who did not have myopia, defined by a mean spherical equivalent $\geq$ -0.5D.

A Receiver Operating Characteristic (ROC) curve was drawn for each case and the Area Under the Curve (AUC) was calculated. R programming language and software environment for statistical computing (https://cran.r-project.org/) was used for both the logistic regression models ('glm') and to evaluate the performance of the model ('ROCR').

Polymorphisms associated at a GWAS level ($P<5x10^{-08}$) were clustered within an "associated genomic region", defined as a contiguous genomic region where GWAS-significant markers were within 1 million base pairs from each other. Significant polymorphisms were annotated with the gene inside whose transcript-coding region they are located, or alternatively, if located between two genes, with the gene nearest to it. The associated genomic regions were collectively annotated with the gene overlapping, or nearest the most significantly associated variant within that region.

The known relationships between identified genetic loci and other phenotypic traits were derived from two datasets: the Online Mendelian Inheritance In Man (OMIM, https://omim.org), which is a

continuously curated catalog of human genes and phenotypic changes their polymorphic forms cause in humans and the GWAS Catalog[80] which is a curated catalog of previous GWAS association of SNPs or genes with other phenotypic traits.

The R (https://cran.r-project.org) package MendelianRandomization v3.4.4 was used for Mendelian randomization analyses.

Disease-relevant tissues and cell types were identified by analyzing gene expression data together with summary statistics from the meta-analysis of refractive error in all five cohorts, as described elsewhere[81]. Expression data was obtained from the following sources: 1) the GTEx release v7 (https://gtexportal.org/home/datasets) 2) Fetal and adult corneal, trabecular meshwork and ciliary body RNA sequencing data previously described [82] and 3) data from the subset of subjects with presumed healthy adult retinas (AMD=1) from datasets described elsewhere[83].

As the transcription data were heterogeneous and in different units, expression levels for all tissues were rank-transformed. Hierarchical clustering was used to help visualize similarities and differences of patterns of transcript expression across different tissues ('hclust' package in R).

SMR (Summary data–based Mendelian randomization) uses GWAS variants as instrumental variables and gene expression levels or methylation levels as mediating traits, in order to test whether the causal effect of a specific variant on the phenotype-of-interest acts via a specific gene[84]. The SMR tests were performed used three different: the summary statistics of eQTL associations in the untransformed peripheral blood samples of 5,311 subjects[85], as well as eQTL effects and *cis*- methylation analysis (*cis*-mQTL), both in brain tissues[86].

The Gene-Set Enrichment Analysis (GSEA) was implemented in the MAGENTA software[87]. We used the versions from September 2017.

Results of three statistical tests for natural selection were imported from the 1000 Genomes Selection Browser[88].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Summary statistics from the cohorts participating in the meta-analysis can be downloaded from ftp://twinr-ftp.kcl.ac.uk/Refractive_Error_MetaAnalysis_2020/ and public repositories such as the GWAS Catalogue (https://www.ebi.ac.uk/gwas/downloads/summary-statistics). These freely downloadable summary statistics are calculated using all cohorts described in this manuscript, except for the 23andMe participants. This is due to a non-negotiable clause in the 23andMe data transfer agreement, intended to protect the privacy of the 23andMe research participants.

To fully recreate our meta-analytic results, all bona fide researchers can obtain the 23andMe summary statistics by emailing 23andMe (dataset-request@23andme.com) and subsequently meta-analyzing them along the freely accessible summary statistics for all the other cohorts.

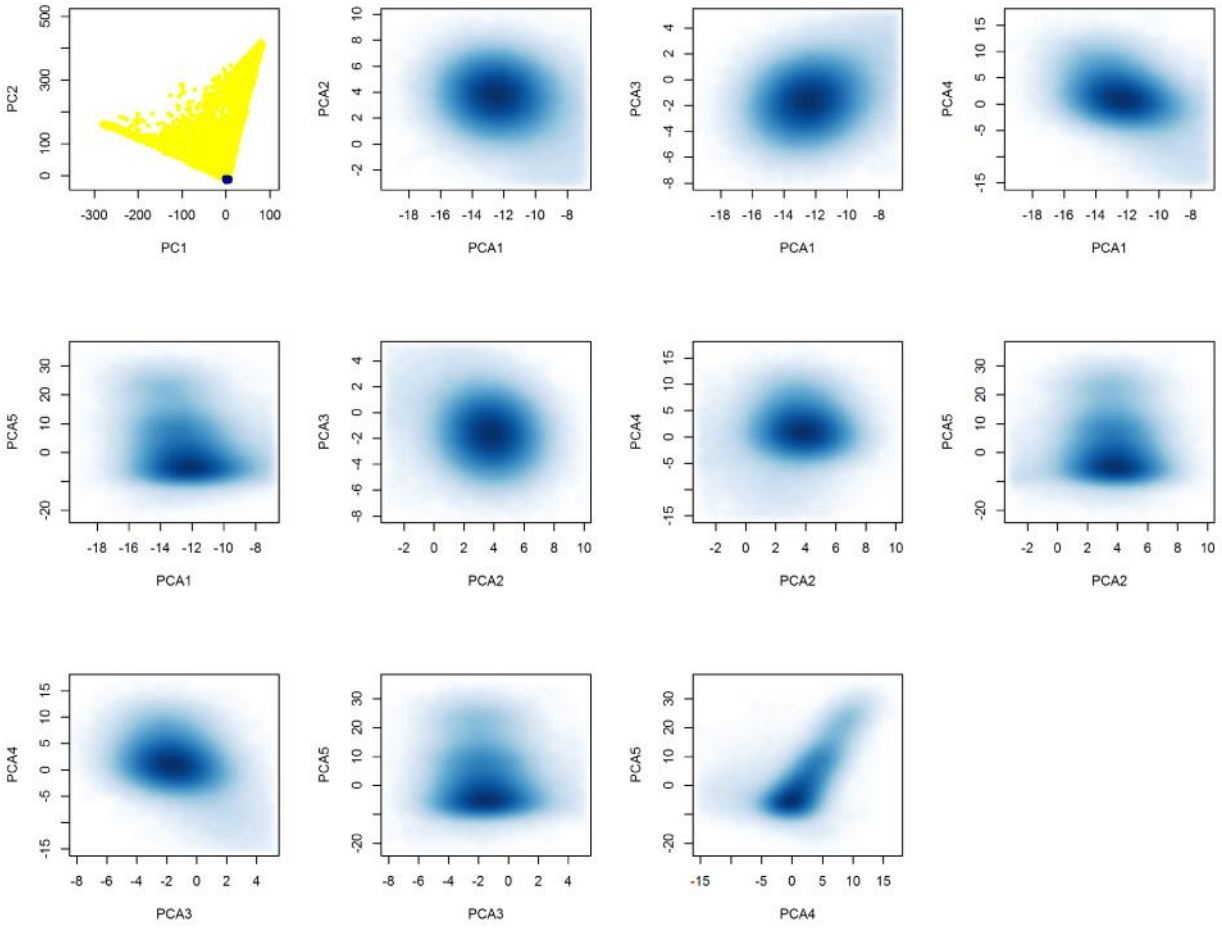
References

64. Cumberland, P.M. *et al.* Frequency and Distribution of Refractive Error in Adult Life: Methodology and Findings of the UK Biobank Study. *PLoS One* **10**, e0139780 (2015).
65. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203-209 (2018).
66. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2011).
67. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P. & Price, A.L. Mixed-model association for biobank-scale datasets. *Nature genetics*, 1 (2018).
68. Fuchsberger, C., Abecasis, G.R. & Hinds, D.A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782-4 (2015).
69. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285-95 (2015).
70. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
71. Tedja, M.S. *et al.* Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. *Nat Genet* **50**, 834-848 (2018).
72. Riboli, E. & Kaaks, R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* **26 Suppl 1**, S6-14 (1997).
73. Hayat, S.A. *et al.* Cohort profile: A prospective cohort study of objective physical and cognitive capability and visual health in an ageing population of men and women in Norfolk (EPIC-Norfolk 3). *Int J Epidemiol* **43**, 1063-72 (2014).
74. Delaneau, O., Marchini, J., Genomes Project, C. & Genomes Project, C. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**, 3934 (2014).
75. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
76. Winkler, T.W. *et al.* Quality control and conduct of genome-wide association meta-analyses. *Nat Protoc* **9**, 1192-212 (2014).
77. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
78. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
79. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
80. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
81. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).
82. Carnes, M.U., Allingham, R.R., Ashley-Koch, A. & Hauser, M.A. Transcriptome analysis of adult and fetal trabecular meshwork, cornea, and ciliary body tissues by RNA sequencing. *Exp Eye Res* **167**, 91-99 (2018).

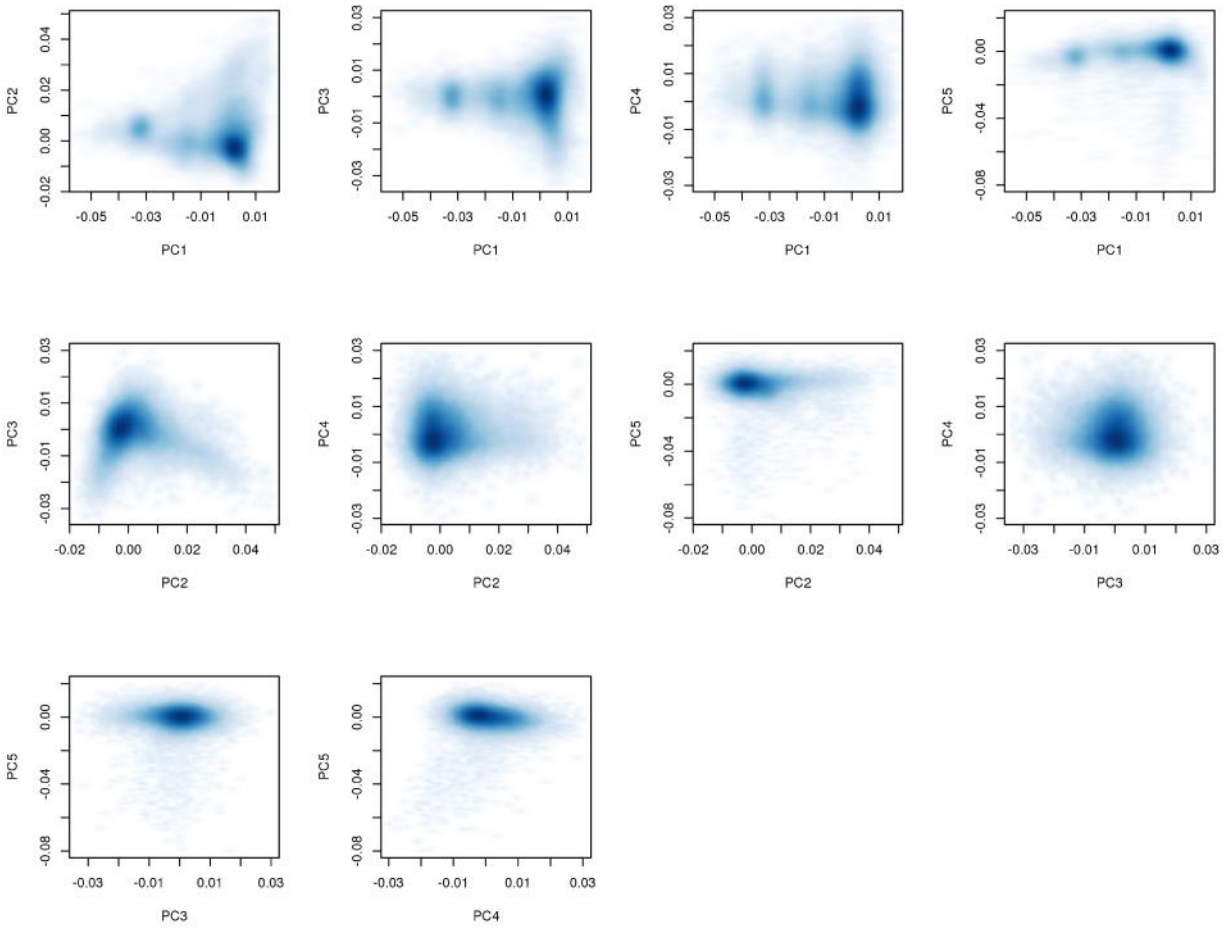
83. Ratnapriya, R. *et al.* Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat Genet* **51**, 606-610 (2019).
84. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
85. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-1243 (2013).
86. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun* **9**, 2282 (2018).
87. Segre, A.V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6**(2010).
88. Pybus, M. *et al.* 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res* **42**, D903-9 (2014).

Supplementary Figure 1. Principal component plots of the subjects in each of the participating cohorts.

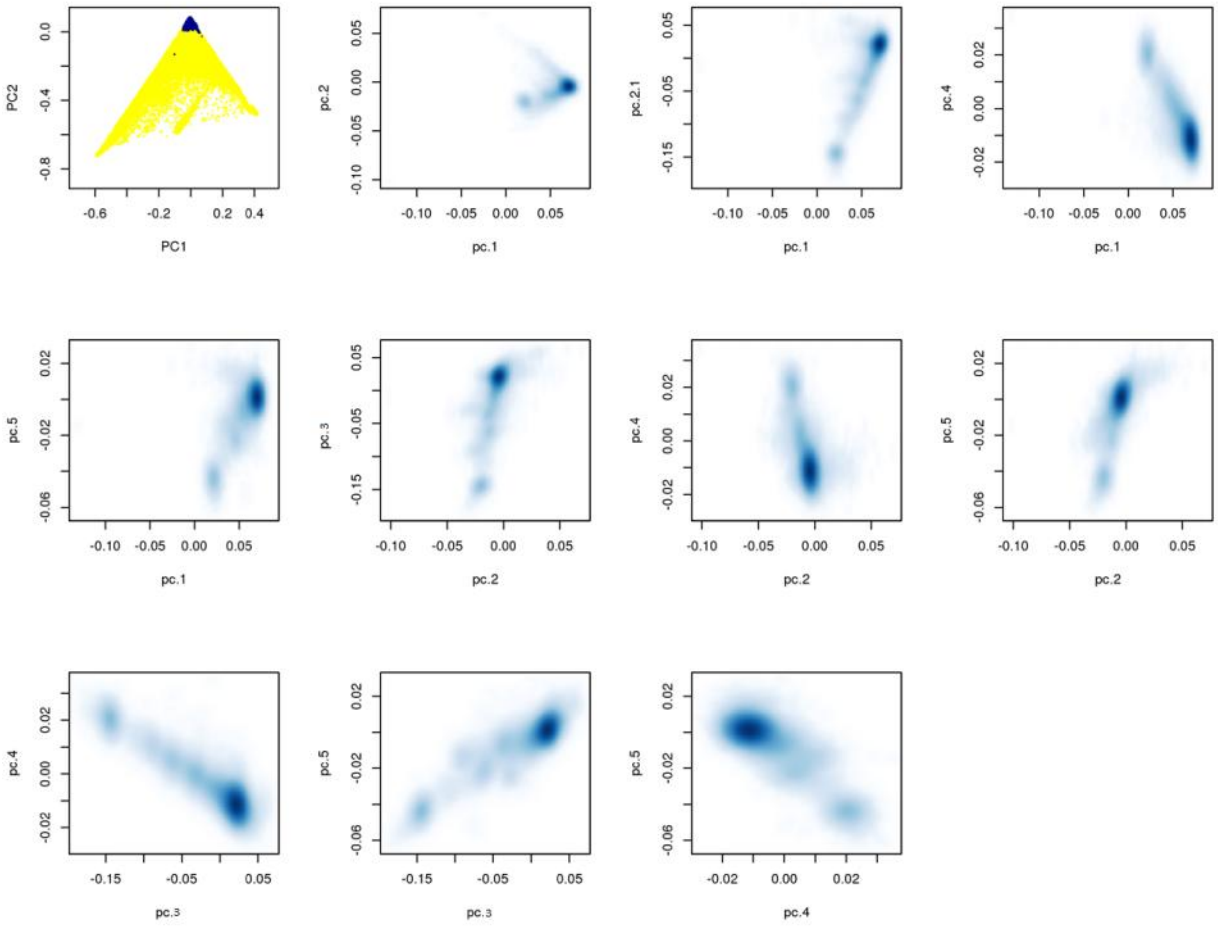
A. UK Biobank



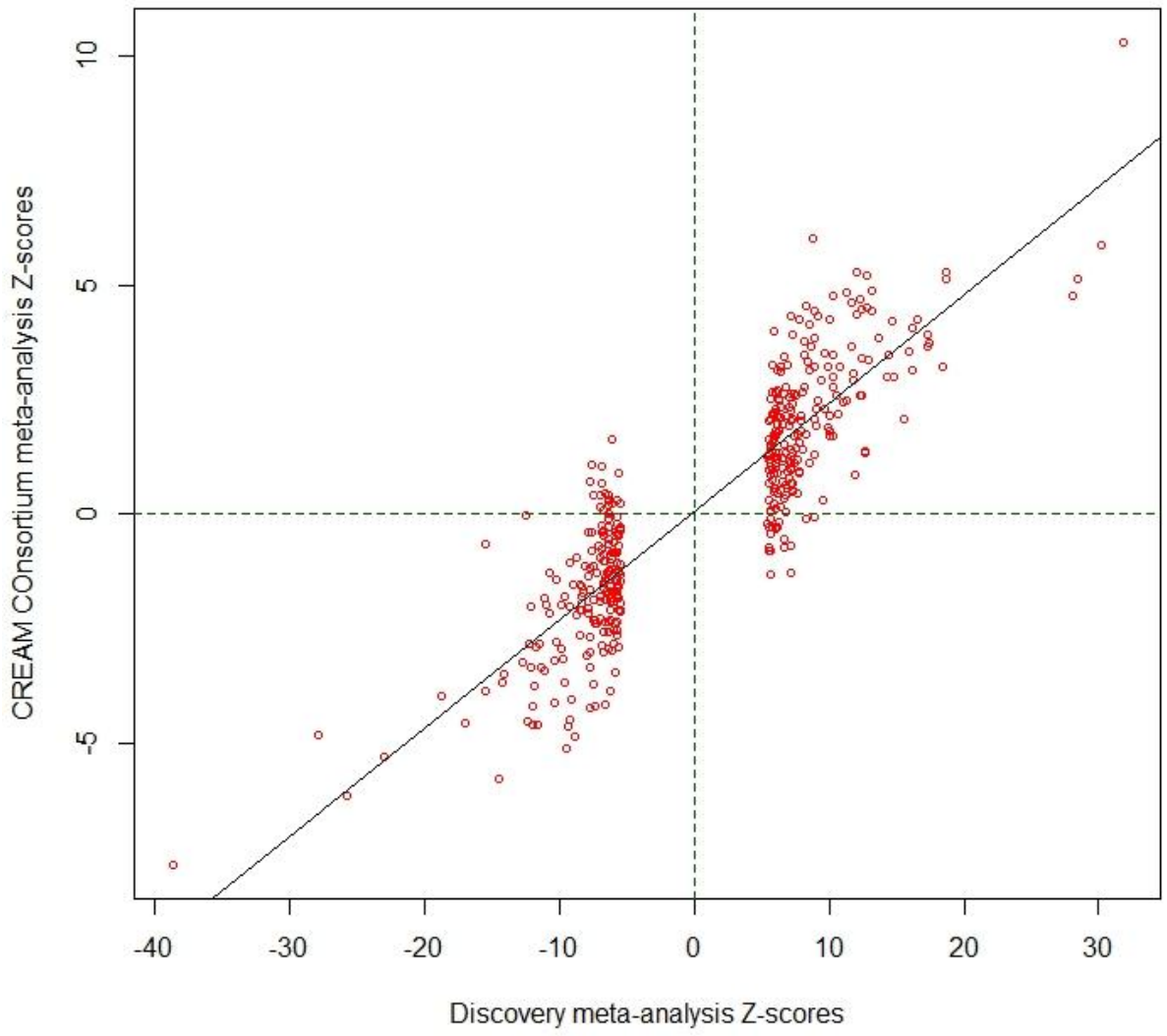
B. GERA


PC2
PC1
PC3
PC1
PC4
PC1
PC5
PC1
PC3
PC2
PC4
PC2
PC5
PC2
PC4
PC3
PC5
PC3
PC5
PC4

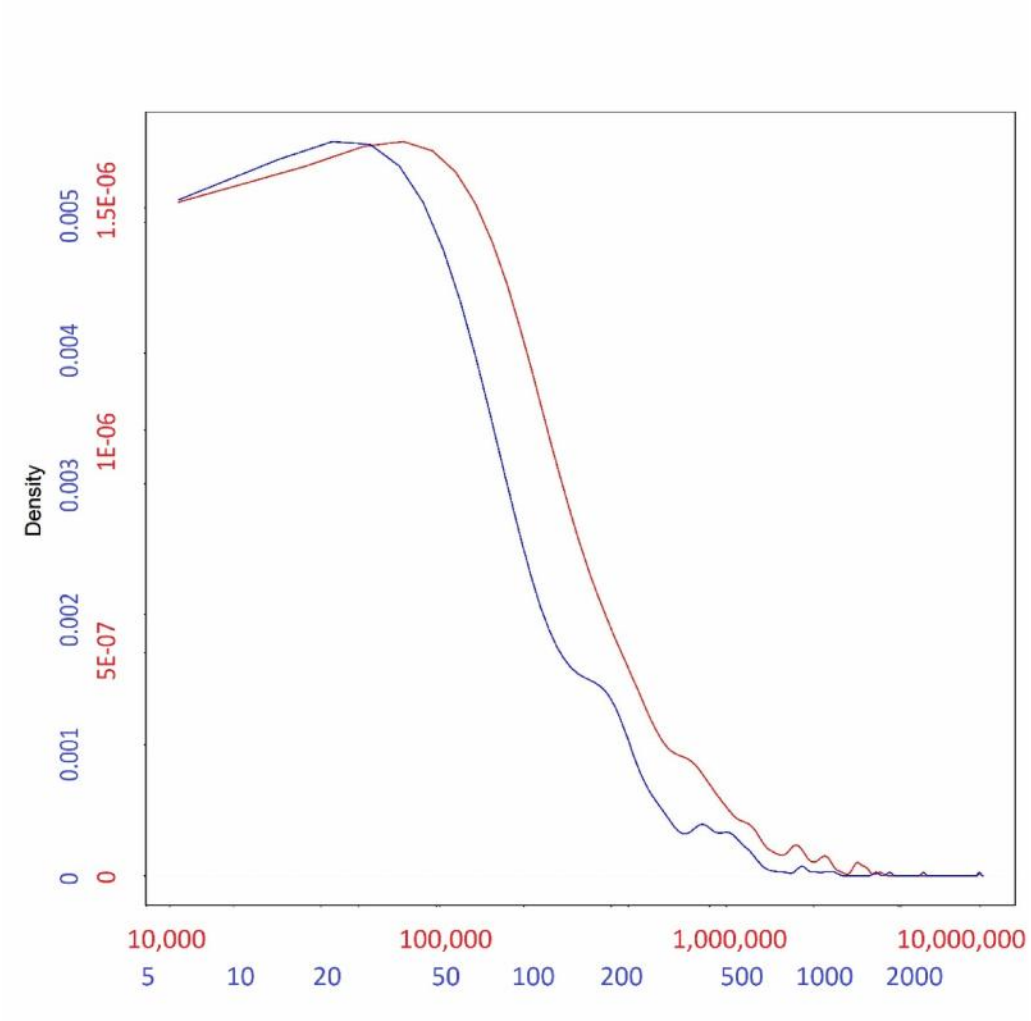
C. 23andMe


PC2
PC1
pc.2
pc.1
pc.2.1
pc.1
pc.4
pc.1
pc.5
pc.1
pc.3
pc.2
pc.4
pc.2
pc.5
pc.2
pc.4
pc.3
pc.5
pc.3
pc.5
pc.4

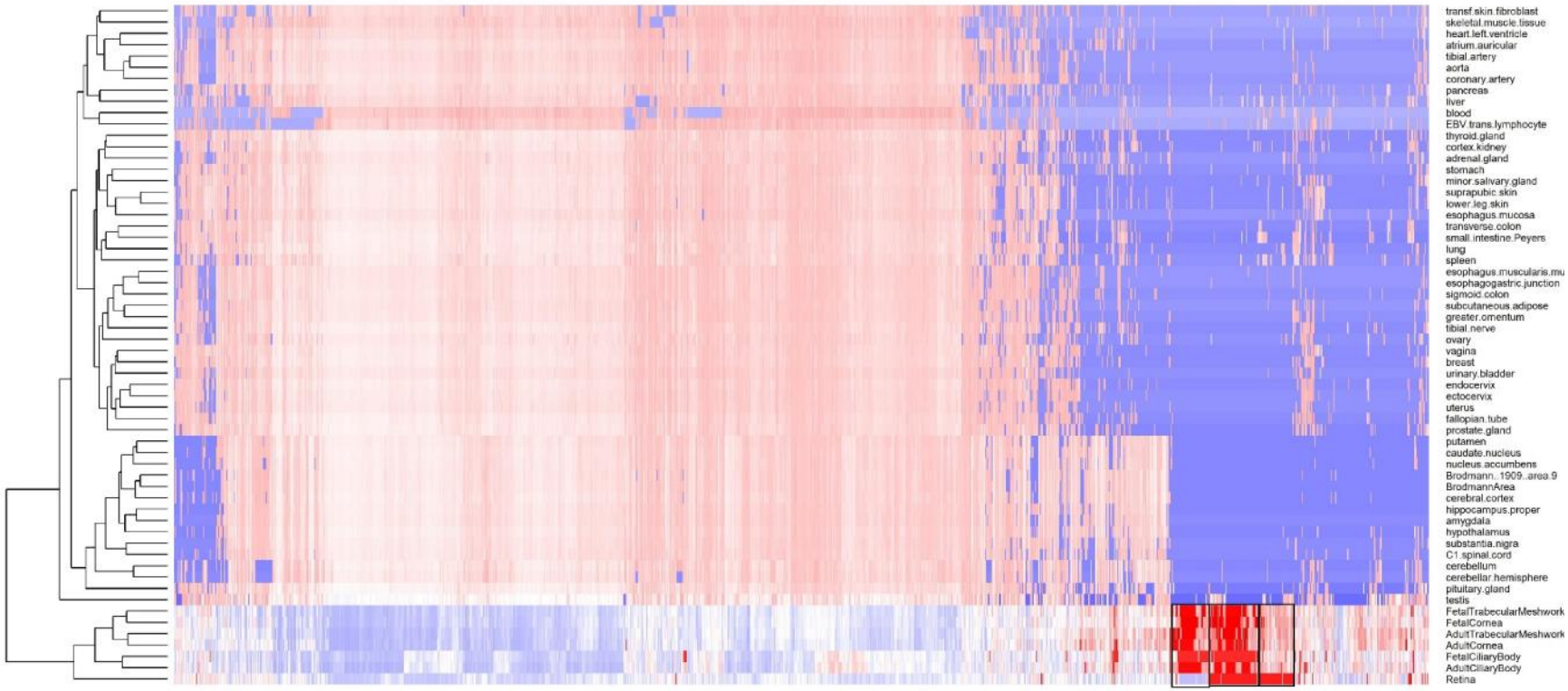
Supplementary Figure 2. Correlation of effect sizes between the discovery cohort meta-analysis (UK Biobank analysis on spherical equivalent + GERA, spherical equivalent + 23andMe, self-reported myopia cases and controls + UK Biobank inferred myopia cases and controls) and those from the CREAM Consortium participants, used as replication. The z-scores for the discovery are on the y-axis and those from the CREAM cohort in the x-axis.



Supplementary Figure 3. Distribution of the base-pair length (red) of the 449 regions associated in the meta-analysis of all available cohorts (Supplementary Table 3), alongside the distribution of number of SNPs (blue) for each region. Numbers in each of the axes in the figure are differentially colored to match the density curve they correspond to: red for the length of the region and blue for the number of SNPs.

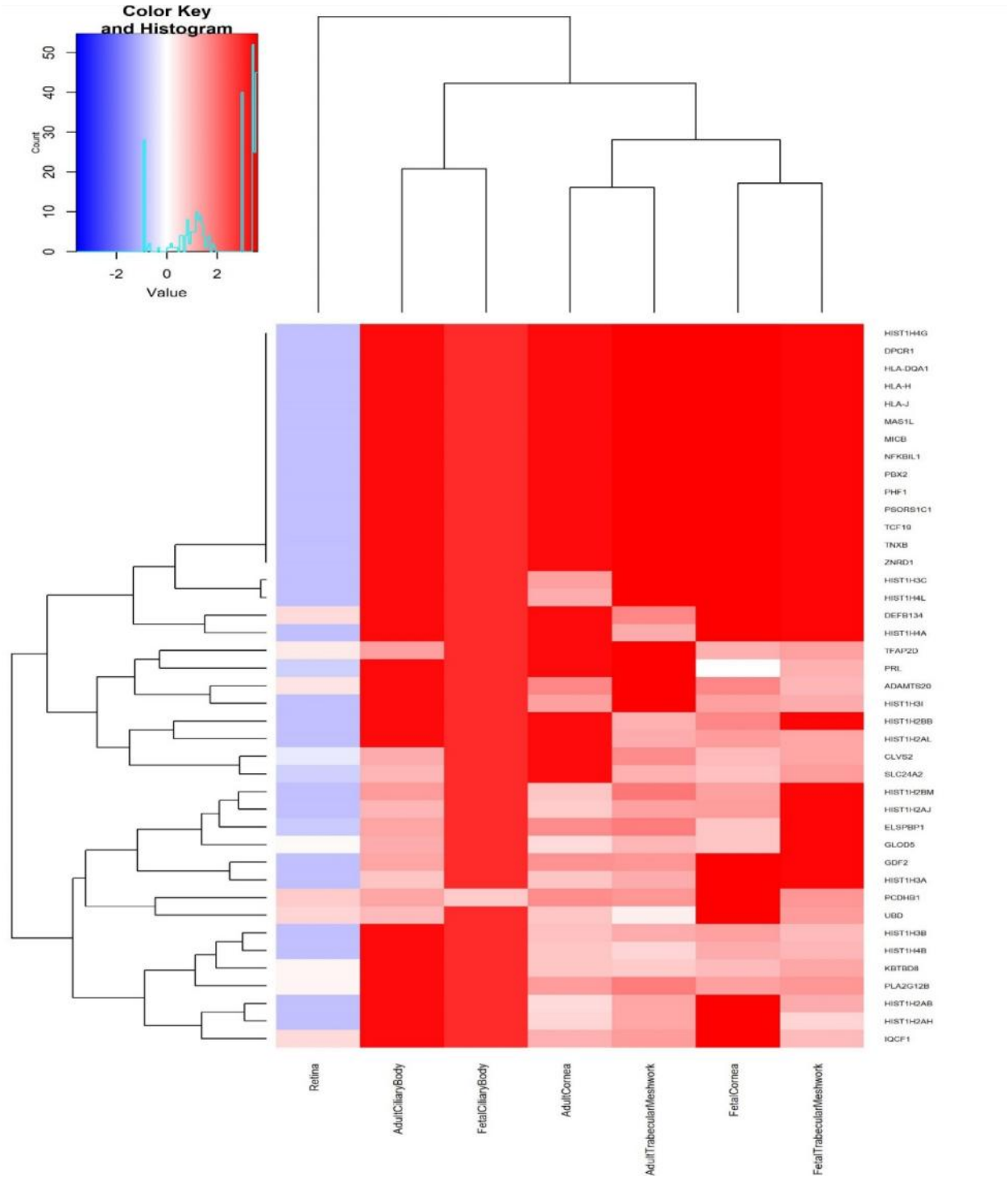


Supplementary Figure 4. Expression of genes located in the associated loci (Supplementary Table 3) along the x-axis, across several human body tissues (y-axis). The colors represent the centile ranking of the expression level of the gene in the tissue of interest. The hotter colors represent higher ranking of the gene expression and the colder colors low expression. Both genes and tissues are clustered in accordance with their pattern similarity. The symbol of all the genes could not be visualized and therefore are removed for the sake of clarity. Eye tissues, whether fetal or adult, appear to have similar patterns of gene expressions (clustered together at the bottom of the figure). Genes that are highly expressed in eye tissues fall in three clusters, shown with a black box. These clusters are displayed in more detail in Figure 4A, B and C.

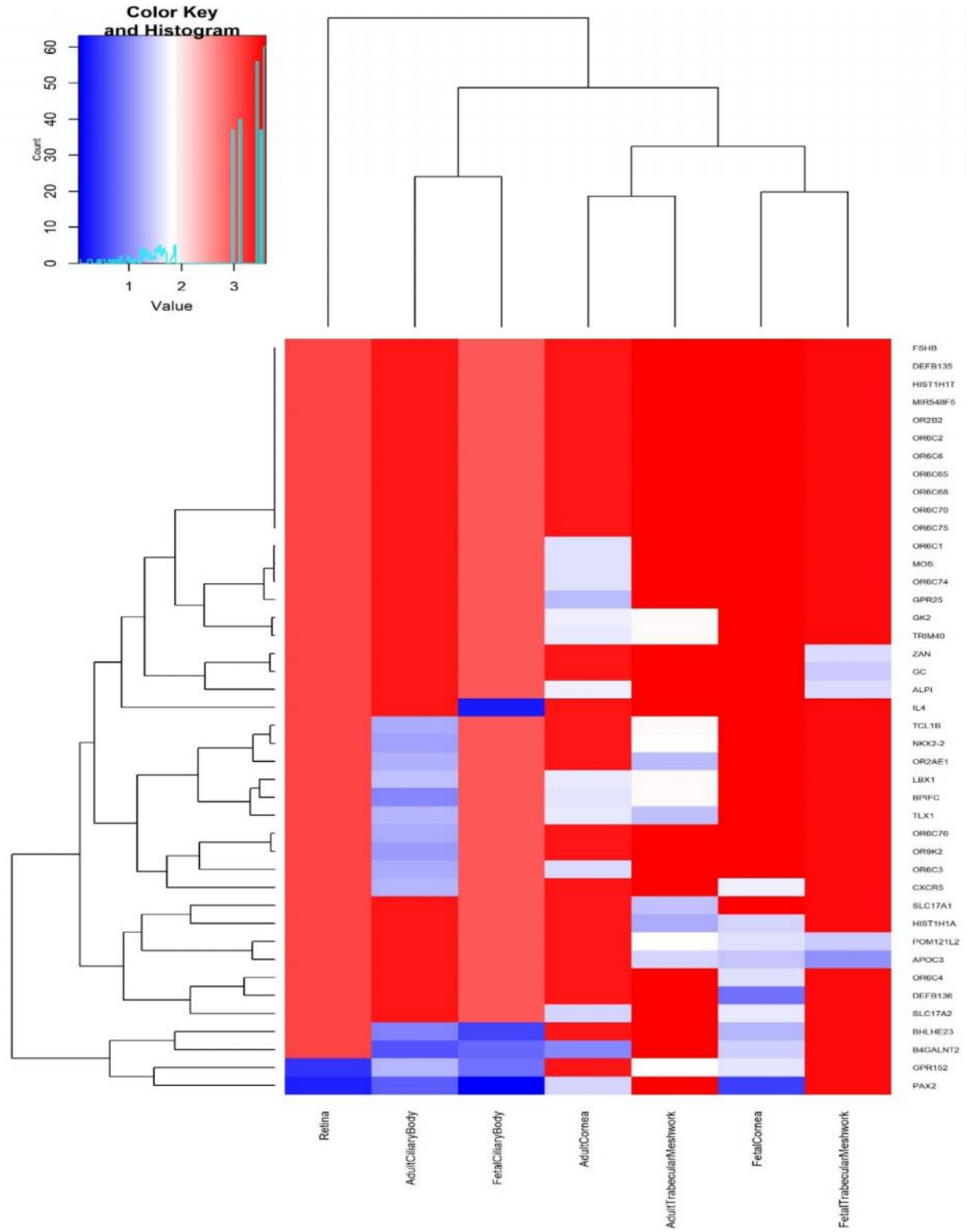


Supplementary Figure 5. Genes from the regions associated with RE (from Supplementary Table 3) that are particularly expressed in eye tissues, compared to non-ocular tissues. These clusters are those highlighted in Supplementary Figure 3, but for the sake of clarity they are shown in transposed orientation compared to the previous figure (here genes in the y-axis and eye tissues in the x-axis), but same color codes as before. The dendrograms represent the degree of similarity observed for both tissues and gene expressions. The clusters are given in the order in which they were clustered together, from left to right: A) genes that are expressed more in other ocular tissues (fetal and adult) but much less in the adult retina. B) genes that are highly expressed in the retina and other ocular tissues, and C) genes that are expressed in the retina, but less in the other ocular tissues tested.

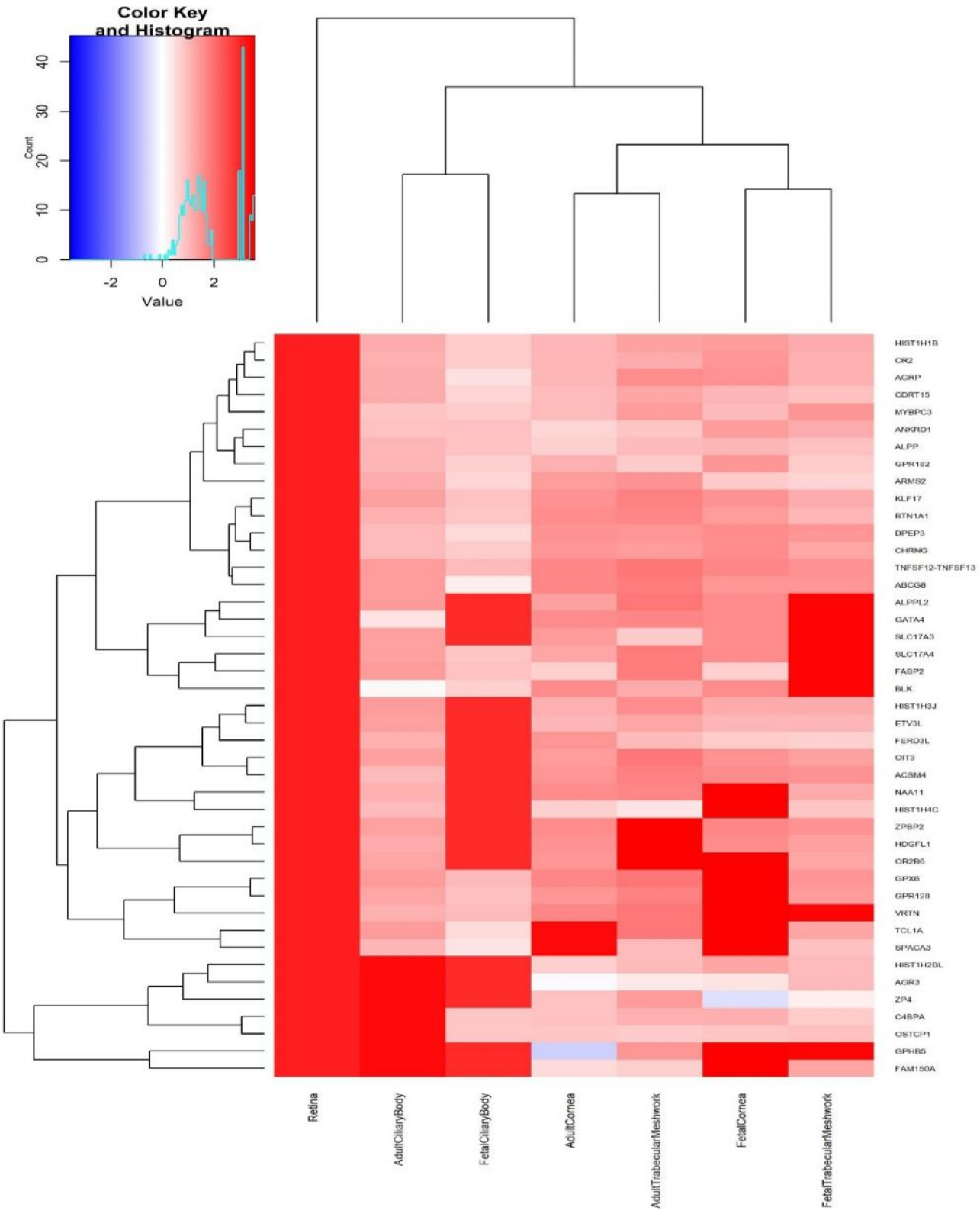
A)



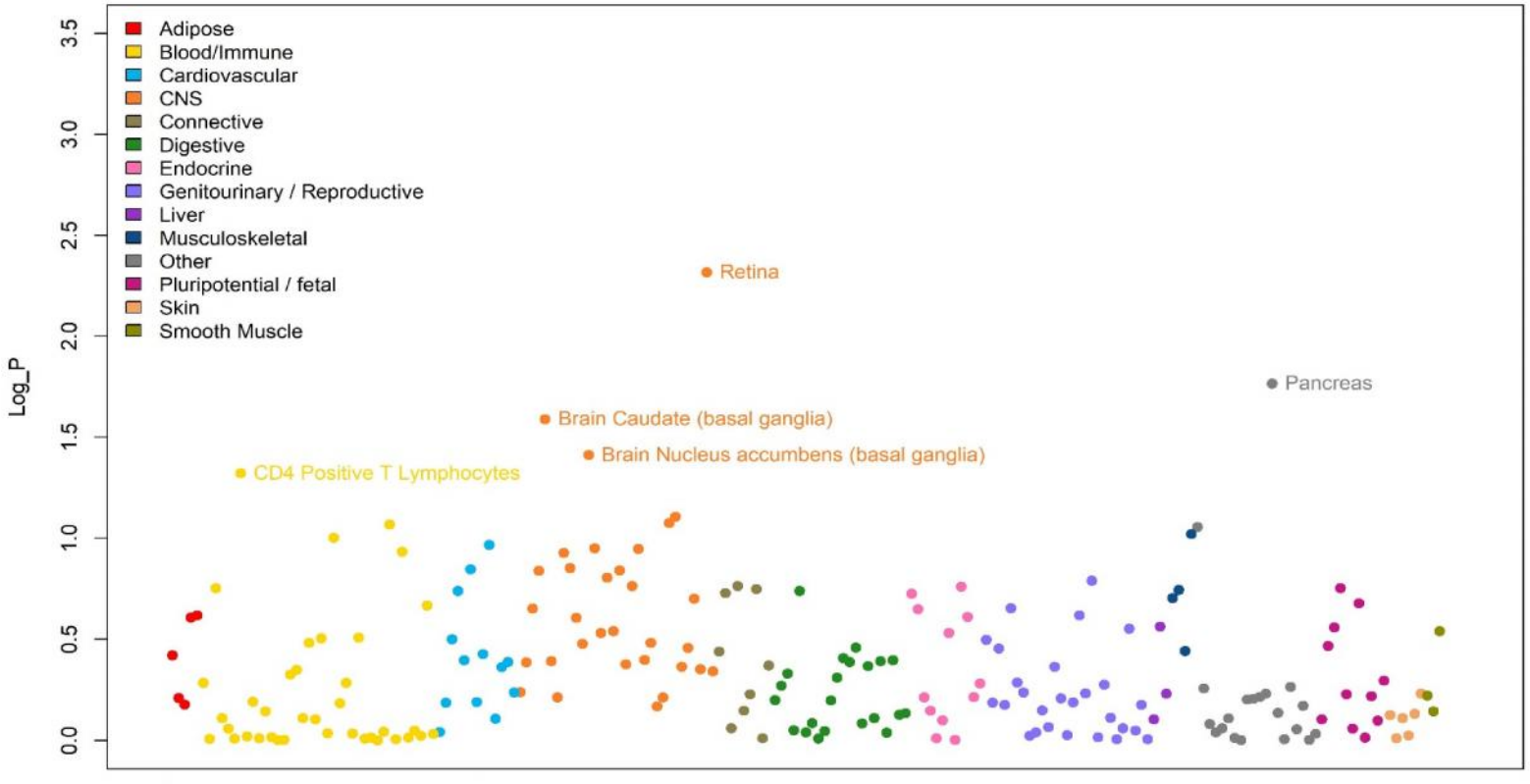
B)


Color Key and Histogram
Count
Value
FSHB
DEFB135
HIST1H1T
MIR548F5
OR2B2
OR6C2
OR6C6
OR6C65
OR6C68
OR6C70
OR6C75
OR6C1
MOS
OR6C74
GPR25
GK2
TRIM40
ZAN
GC
ALPI
IL4
TCL1B
NKX2-2
OR2AE1
LBX1
BPIFC
TLX1
OR6C76
OR9K2
OR6C3
CXCR5
SLC17A1
HIST1H1A
POM121L2
APOC3
OR6C4
DEFB136
SLC17A2
BHLHE23
B4GALNT2
GPR152
PAX2
Retina
AdultCiliaryBody
FetalCiliaryBody
AdultCornea
AdultTrabecularMeshwork
FetalCornea
FetalTrabecularMeshwork

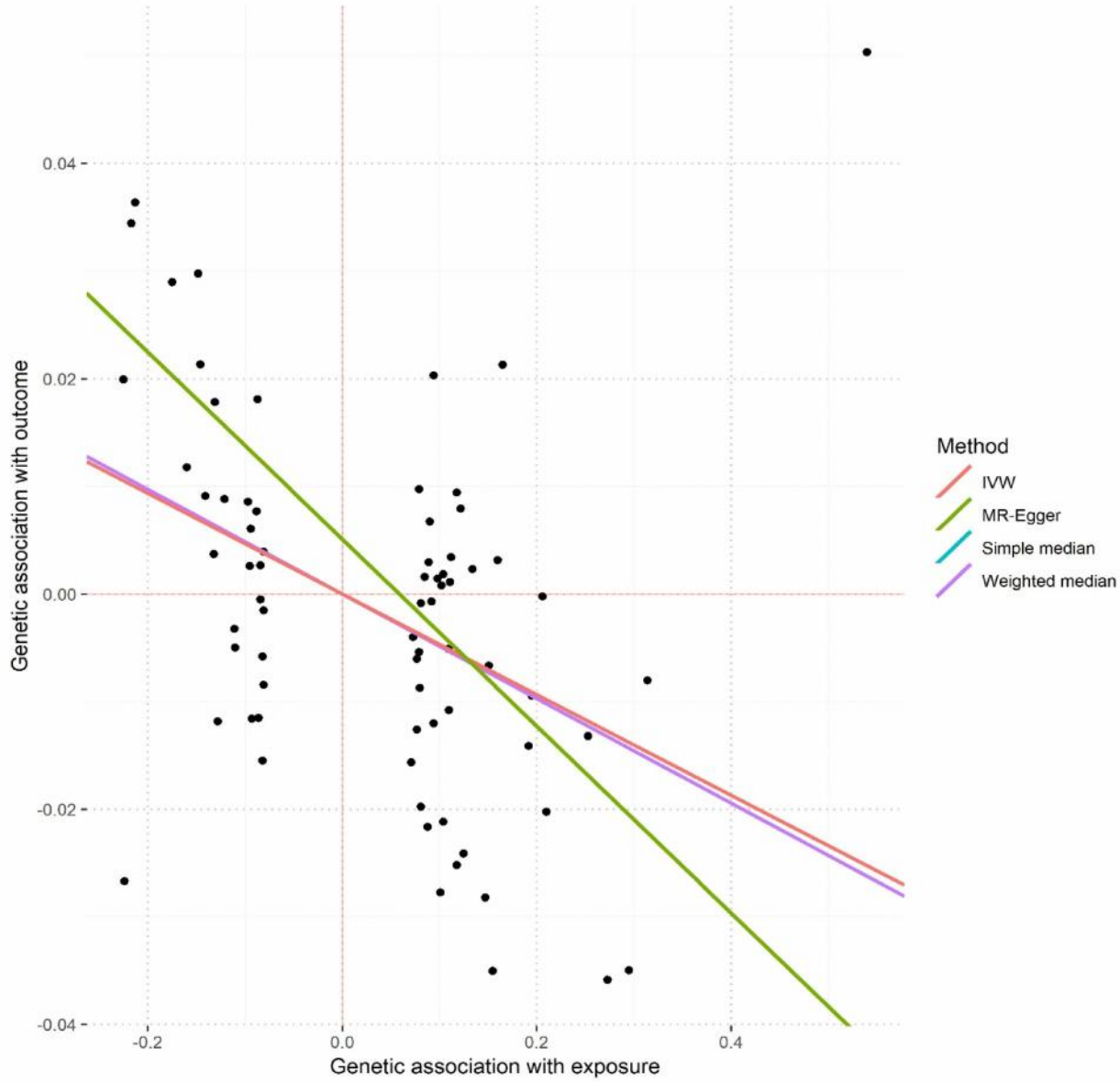
C)


Color Key and Histogram
Count
0
10
20
30
40
-2
0
2
Value
HIST1H1B
CR2
AGRP
CDRT15
MYBPC3
ANKRD1
ALPP
GPR182
ARMS2
KLF17
BTN1A1
DPEP3
CHRNG
TNFSF12-TNFSF13
ABCG8
ALPPL2
GATA4
SLC17A3
SLC17A4
FABP2
BLK
HIST1H3J
ETV3L
FERD3L
OIT3
ACSM4
NAA11
HIST1H4C
ZPBP2
HDGFL1
OR2B6
GPX6
GPR128
VRTN
TCL1A
SPACA3
HIST1H2BL
AGR3
ZP4
C4BPA
OSTCP1
GPHB5
FAM150A
Retina
AdultCiliaryBody
FetalCiliaryBody
AdultCornea
AdultTrabecularMeshwork
FetalCornea
FetalTrabecularMeshwork

Supplementary Figure 6. Results of the LD score regression analysis applied to specifically expressed genes (LDSC-SEG) on multiple tissue for the meta-analysis results. Each point represents one tissue or cell line (along the x-axis) and the log10 value of the p-value of the correlation between the meta-analysis results and gene expression. There were 205 tests carried out, one in each tissue and cell line, therefore only tissues with a correlation p-value< 0.00025 (Log_P> 3.6 in this figure), would have been significant after multiple testing. This condition was not fulfilled for any of the available tissues.

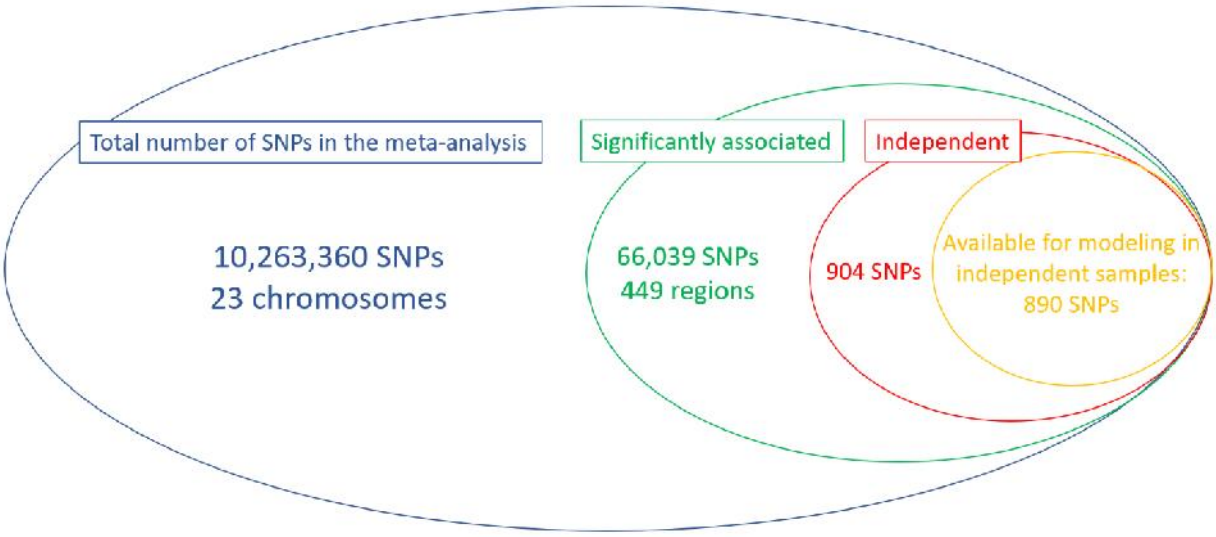


Supplementary Figure 7. Mendelian randomization results on causality of IOP over refractive error. Single points in the graph represent coordinates determined by the effect of each specific SNP over IOP (x-axis, mmHg) and spherical equivalent (y-axis, Diopter units). The lines represent the regression lines from each model, as specified in the figure legend. In some cases, these lines may not visible because they overlap (please refer to the values underneath the figure)

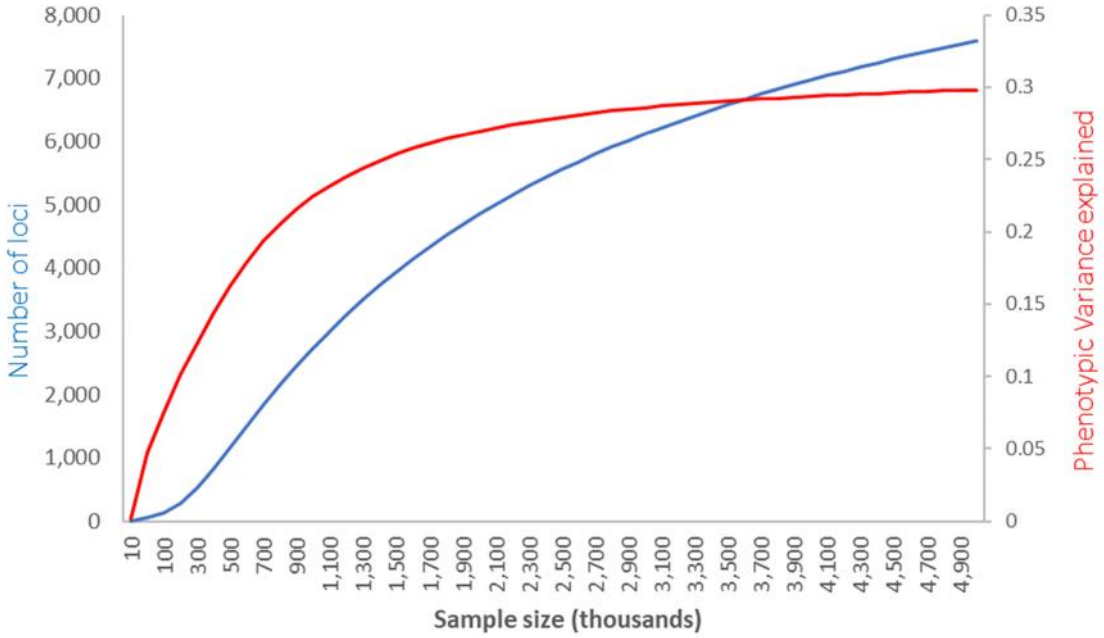


Method	Estimate	SE	95% CI		P-value
Simple median	-0.049	0.02	-0.088	-0.009	0.016
Weighted median	-0.049	0.02	-0.088	-0.009	0.017
IVW	-0.047	0.014	-0.073	-0.02	0.001
MR-Egger	-0.087	0.038	-0.161	-0.013	0.022
(intercept)	0.005	0.004	-0.004	0.014	0.256

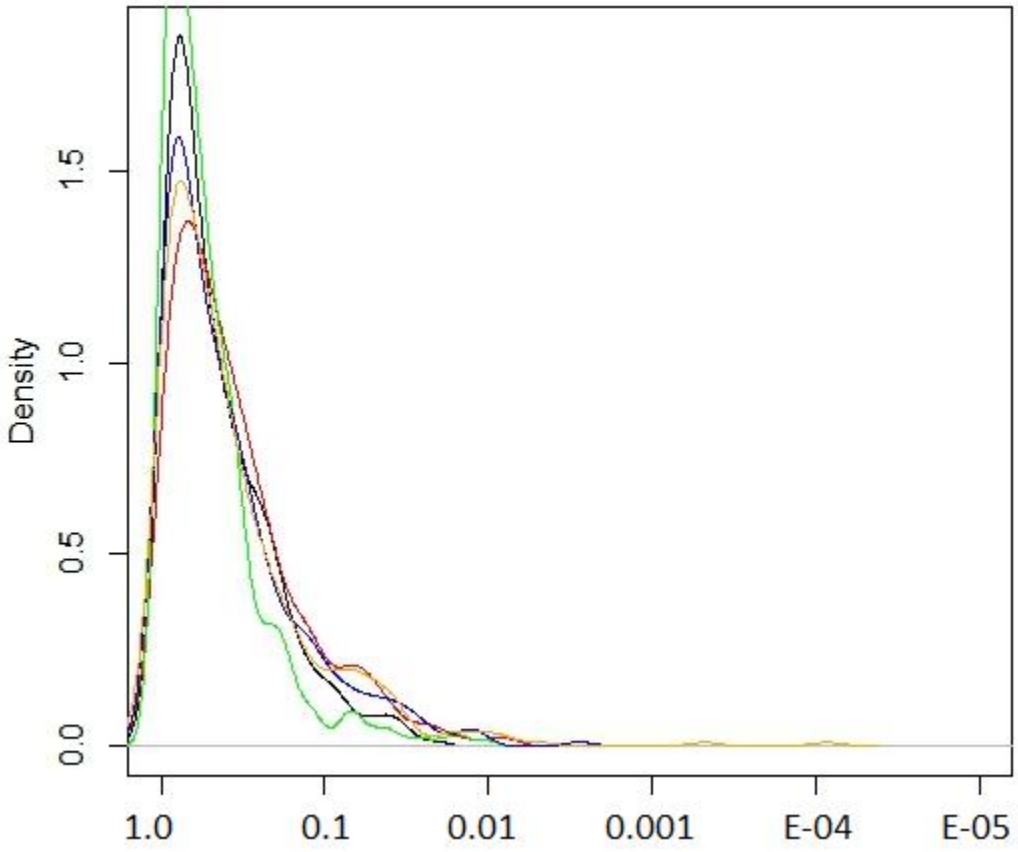
Supplementary Figure 8. Venn's Diagram of the number of SNPs considered in each of the stages of this study. The different circles represent various stages, inclusion in the meta-analysis (blue), identification of significant loci (green), conditional analysis results identifying independent effects (red) and the total number of SNPs available for inclusion in prediction and heritability estimation in the independent (i.e. not part of the original meta-analysis) EPIC-Norfolk cohort (orange).



Supplementary Figure 9. Prediction for the total number of SNPs and phenotypic variance explained as a function of GWAS sample size in future studies, based on the distribution of effects observed in the current meta-analysis. The plot lines show the predicted relationship between the number of loci associated with refractive error (left vertical axis, blue line) and the variance they help explain (red line, right vertical axis), as a function of the sample size (x-axis) used in future GWAS or meta-analyses. These projections are consistent with the observed results, where an effective sample of 379,227 identified 904 independent signals after a conditional analysis, explaining 12-16% of refractive error variability.



Supplementary Figure 10. The distribution of various natural selection test scores for SNPs associated with refractive error. The values on the x-axis represent the ranking in terms of natural selection observed and the y-axis the density of that rank. The different tests show are iHS, XP-EHH (CEU vs YRI), XP-EHH average score, XP-EHH maximum score and Tajima scores (black, green, red, blue and yellow respectively)



Meta-analysis of 542,934 subjects of European ancestry identifies new genes and mechanisms predisposing to refractive error and myopia

Supplementary Notes

Table of Contents

1 UK Biobank Eye and Vision Consortium Membership

- Prof **Tariq ASLAM** - Manchester University, Manchester, United Kingdom
- Prof **Sarah BARMAN** - Kingston University, London, United Kingdom
- Prof **Jenny BARRETT** - University of Leeds, Yorkshire, United Kingdom
- Prof **Paul BISHOP** - Manchester University, Manchester, United Kingdom
- Mr **Peter BLOWS** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Dr **Catey BUNCE** - King's College London, London, United Kingdom
- Dr **Roxana CARARE** - University of Southampton, Southampton, United Kingdom
- Prof **Usha CHAKRAVARTHY** - Queens University Belfast, Belfast, Ireland
- Miss **Michelle CHAN** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Dr **Sharon CHUA** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **David CRABB** – City, University of London, London, United Kingdom
- Mrs **Philippa CUMBERLAND** - UCL Great Ormond Street Institute of Child Health, London, United Kingdom
- Dr **Alexander DAY** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Miss **Parul DESAI** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **Bal DHILLON** - University of Edinburgh, Scotland, United Kingdom
- Prof Andrew DICK - University of Bristol, Bristol, United Kingdom
- Dr **Cathy EGAN** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **Sarah ENNIS** - University of Southampton, Southampton, United Kingdom
- Prof **Paul FOSTER** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Dr **Marcus FRUTTIGER** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Dr **John GALLACHER** - University of Oxford, Oxford, United Kingdom
- Prof **David (Ted) GARWAY-HEATH** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Dr **Jane GIBSON** - University of Southampton, Southampton, United Kingdom
- Mr **Dan GORE** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **Jeremy GUGGENHEIM** - Cardiff University, Wales, United Kingdom
- Prof **Chris HAMMOND** - King's College London, London, United Kingdom
- Prof **Alison HARDCASTLE** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **Simon HARDING** - University of Liverpool, London, United Kingdom

- Dr **Ruth HOGG** - Queens University Belfast, Belfast, Ireland
- Dr **Pirro HYS**I - King's College London, London, United Kingdom
- Mr **Pearse A KEANE** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **Sir Peng Tee KHAW** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Mr **Anthony KHAWAJA** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Mr **Gerassimos LASCARATOS** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **Andrew LOTERY**- University of Southampton, Southampton, United Kingdom
- Prof Phil LUTHERT - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Dr **Tom MACGILLIVRAY** - University of Edinburgh, Scotland, United Kingdom
- Dr **Sarah MACKIE** - University of Leeds, Yorkshire, United Kingdom
- Prof **Keith MARTIN** - University of Cambridge, Cambridge, United Kingdom
- Ms **Michelle MCGAUGHEY** - Queen's University Belfast, Belfast, Ireland
- Dr **Bernadette MCGUINNESS** - Queen's University Belfast, Belfast, Ireland
- Dr **Gareth MCKAY** - Queen's University Belfast, Belfast, Ireland
- Mr **Martin MCKIBBIN** - Leeds Teaching Hospitals NHS Trust, Yorkshire, United Kingdom
- Dr **Danny MITRY** – NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom & Royal Free Hospital, London, United Kingdom
- Prof **Tony MOORE** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **James MORGAN** - Cardiff University, Wales, United Kingdom
- Ms **Zaynah MUTHY** – NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Mr **Eoin O'SULLIVAN** - King's College Hospital NHS Foundation Trust, London, United Kingdom
- Dr **Chris OWEN** - St George's, University of London, London, United Kingdom
- Mr Praveen PATEL - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Mr **Euan PATERSON** - Queens University Belfast, Belfast, Ireland
- Dr **Tunde PETO** - Queen's University Belfast, Belfast, Ireland
- Dr **Axel PETZOLD** - UCL Institute of Neurology, London, United Kingdom
- Prof **Jugnoo RAHI** - UCL Great Ormond Street Institute of Child Health, London, United Kingdom
- Dr **Alicja RUDNICKA** - St George's, University of London, London, United Kingdom
- Mr **Jay SELF** - University of Southampton, Southampton, United Kingdom
- Prof **Sobha SIVAPRASAD** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Mr **David STEEL** - Newcastle University, Newcastle, United Kingdom
- Mrs **Irene STRATTON** - Gloucestershire Hospitals NHS Foundation Trust

- Mr **Nicholas STROUTHIDIS** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **Cathie SUDLOW** - University of Edinburgh, Scotland, United Kingdom
- Dr **Caroline THAUNG** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Miss **Dhanes THOMAS** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Prof **Emanuele TRUCCO** - University of Dundee, Scotland, United Kingdom
- Prof **Adnan TUFAIL** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Dr **Veronique VITART** - University of Edinburgh, Scotland, United Kingdom
- Prof **Stephen VERNON** – Nottingham University Hospitals NHS Trust, Nottingham, United Kingdom
- Mr **Ananth VISWANATHAN** - NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, United Kingdom
- Dr **Cathy WILLIAMS** - University of Bristol, Bristol, United Kingdom
- Dr **Katie WILLIAMS** - King's College London, London, United Kingdom
- Prof **Jayne WOODSIDE** - Queen's University Belfast, Belfast, Ireland
- Dr **Max YATES** - University of East Anglia, Norwich, United Kingdom
- Ms **Jennifer YIP** - University of Cambridge, Cambridge, United Kingdom
- Dr **Yalin ZHENG** - University of Liverpool, London, United Kingdom

2 The CREAM Consortium

Joan E. Bailey-Wilson[1], Paul Nigel Baird[2], Amutha Barathi Veluchamy[3-5], Ginevra Biino[6], Kathryn P. Burdon[7], Harry Campbell[8], Li Jia Chen[9], Ching-Yu Cheng[10-12], Emily Y. Chew[13], Jamie E. Craig[14], Phillippa M. Cumberland[15], Margaret M. Deangelis[16], Cécile Delcourt[17], Xiaohu Ding[18], Cornelia M. van Duijn[19], David M. Evans[20-22], Qiao Fan[23], Maurizio Fossarello[24], Paul J. Foster[25], Puya Gharahkhani[26], Adriana I. Iglesias[19,27,28], Jeremy A. Guggenheim[29], Xiaobo Guo[18,30], Annechien E.G. Haarman[19,28], Toomas Haller[31], Christopher J. Hammond[32], Xikun Han[26], Caroline Hayward[33], Mingguang He[2,18], Alex W. Hewitt[2,7,34], Quan Hoang[3,35] , Pirro G. Hysi[32], Robert P. Igo Jr.[36], Sudha K. Iyengar[36-38], , Jost B. Jonas[39,40], Mika Kähönen[41,42], Jaakko Kaprio[43,44], , Anthony P. Khawaja[25,45], Caroline C. W. Klaver[19,28,46], Barbara E. Klein[47], Ronald Klein[47], Jonathan H. Lass[36,37], Kris Lee[47], Terho Lehtimäki[48,49], Deyana Lewis[1], Qing Li[50], Shi-Ming Li[40], Leo-Pekka Lyytikäinen[48,49], Stuart MacGregor[26], David A. Mackey[2,7,34], Nicholas G. Martin[51], Akira Meguro[52], Andres Metspalu[31], Candace Middlebrooks[1], Masahiro Miyake[53], Nobuhisa Mizuki[52], Anthony Musolf[1], Stefan Nickels[54], Konrad Oexle[55], Chi Pui Pang[9], Olavi Pärssinen[56,57], Andrew D. Paterson[58], Norbert Pfeiffer[54], Ozren Polasek[59,60], Jugnoo S. Rahi[15,25,61], Olli Raitakari[62,63], Igor Rudan[8], Srujana Sahebjada[2], Seang-Mei Saw[64,65], Dwight Stambolian[66], Claire L. Simpson[1,67], E-Shyong Tai[65], Milly S. Tedja[19,28], J. Willem L. Tideman[19,28], Akitaka Tsujikawa[53], Virginie J.M. Verhoeven[19,27,28], Veronique Vitart[33], Ningli Wang[40], Juho Wedenoja[43,68], Wen Bin Wei[69], Cathy Williams[22], Katie M. Williams[32], James F. Wilson[8,33], Robert Wojciechowski[1,70,71], Ya Xing Wang[40], Kenji Yamashiro[72], Jason C. S. Yam[9], Maurice K.H. Yap[73], Seyhan Yazar[34], Shea Ping Yip[74], Terri L. Young[47], Xiangtian Zhou[75]

Affiliations

1. Computational and Statistical Genomics Branch, National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland, USA.
2. Centre for Eye Research Australia, Ophthalmology, Department of Surgery, University of Melbourne, Royal Victorian Eye and Ear Hospital, Melbourne, Australia.
3. Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.
4. Duke-NUS Medical School, Singapore, Singapore.
5. Department of Ophthalmology, National University Health Systems, National University of Singapore, Singapore.
6. Institute of Molecular Genetics, National Research Council of Italy, Pavia, Italy.
7. Department of Ophthalmology, Menzies Institute of Medical Research, University of Tasmania, Hobart, Australia.
8. Centre for Global Health Research, Usher Institute for Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK.
9. Department of Ophthalmology and Visual Sciences, The Chinese University of Hong Kong, Hong Kong Eye Hospital, Kowloon, Hong Kong.
10. Department of Ophthalmology, Yong Loo Lin School of Medicine, National University of Singapore, Singapore.
11. Ocular Epidemiology Research Group, Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.
12. Ophthalmology & Visual Sciences Academic Clinical Program (Eye ACP), Duke-NUS Medical School, Singapore.
13. Division of Epidemiology and Clinical Applications, National Eye Institute/National Institutes of Health, Bethesda, USA.
14. Department of Ophthalmology, Flinders University, Adelaide, Australia.

15. Great Ormond Street Institute of Child Health, University College London, London, UK.
16. Department of Ophthalmology and Visual Sciences, John Moran Eye Center, University of Utah, Salt Lake City, Utah, USA.
17. Université de Bordeaux, Inserm, Bordeaux Population Health Research Center, team LEHA, UMR 1219, F-33000 Bordeaux, France.
18. State Key Laboratory of Ophthalmology, Zhongshan Ophthalmic Center, Sun Yat-sen University, Guangzhou, China.
19. Department of Epidemiology, Erasmus Medical Center, Rotterdam, The Netherlands.
20. Translational Research Institute, University of Queensland Diamantina Institute, Brisbane, Queensland, Australia.
21. MRC Integrative Epidemiology Unit, University of Bristol, Bristol, UK.
22. Department of Population Health Sciences, Bristol Medical School, Bristol, UK.
23. Centre for Quantitative Medicine, DUKE-National University of Singapore, Singapore.
24. University Hospital 'San Giovanni di Dio', Cagliari, Italy.
25. NIHR Biomedical Research Centre, Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology, London, UK.
26. Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Australia.
27. Department of Clinical Genetics, Erasmus Medical Center, Rotterdam, The Netherlands.
28. Department of Ophthalmology, Erasmus Medical Center, Rotterdam, The Netherlands.
29. School of Optometry & Vision Sciences, Cardiff University, Cardiff, UK.
30. Department of Statistical Science, School of Mathematics, Sun Yat-Sen University, Guangzhou, China.
31. Institute of Genomics, University of Tartu, Tartu, Estonia.
32. Section of Academic Ophthalmology, School of Life Course Sciences, King's College London, London, UK.
33. MRC Human Genetics Unit, MRC Institute of Genetics & Molecular Medicine, University of Edinburgh, Edinburgh, UK.
34. Centre for Ophthalmology and Visual Science, Lions Eye Institute, University of Western Australia, Perth, Australia.
35. Department of Ophthalmology, Columbia University, New York, USA.
36. Department of Population and Quantitative Health Sciences, Case Western Reserve University, Cleveland, Ohio, USA.
37. Department of Ophthalmology and Visual Sciences, Case Western Reserve University and University Hospitals Eye Institute, Cleveland, Ohio, USA.
38. Department of Genetics, Case Western Reserve University, Cleveland, Ohio, USA.
39. Department of Ophthalmology, Medical Faculty Mannheim of the Ruprecht-Karls-University of Heidelberg, Mannheim, Germany.
40. Beijing Tongren Eye Center, Beijing Tongren Hospital, Beijing Institute of Ophthalmology, Beijing Key Laboratory of Ophthalmology and Visual Sciences, Capital Medical University, Beijing, China.
41. Department of Clinical Physiology, Tampere University Hospital and School of Medicine, University of Tampere, Tampere, Finland.
42. Finnish Cardiovascular Research Center, Faculty of Medicine and Life Sciences, University of Tampere, Tampere, Finland.
43. Department of Public Health, University of Helsinki, Helsinki, Finland.
44. Institute for Molecular Medicine Finland FIMM, HiLIFE Unit, University of Helsinki, Helsinki, Finland.
45. Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK.
46. Department of Ophthalmology, Radboud University Medical Center, Nijmegen, The Netherlands.
47. Department of Ophthalmology and Visual Sciences, University of Wisconsin–Madison, Madison, Wisconsin, USA.

48. Department of Clinical Chemistry, Finnish Cardiovascular Research Center-Tampere, Faculty of Medicine and Life Sciences, University of Tampere.
49. Department of Clinical Chemistry, Fimlab Laboratories, University of Tampere, Tampere, Finland.
50. National Human Genome Research Institute, National Institutes of Health, Baltimore, USA.
51. Genetic Epidemiology, QIMR Berghofer Medical Research Institute, Brisbane, Australia.
52. Department of Ophthalmology, Yokohama City University School of Medicine, Yokohama, Kanagawa, Japan.
53. Department of Ophthalmology and Visual Sciences, Kyoto University Graduate School of Medicine, Kyoto, Japan.
54. Department of Ophthalmology, University Medical Center of the Johannes Gutenberg University Mainz, Mainz, Germany.
55. Institute of Neurogenomics, Helmholtz Zentrum München, German Research Centre for Environmental Health, Neuherberg, Germany.
56. Department of Ophthalmology, Central Hospital of Central Finland, Jyväskylä, Finland.
57. Gerontology Research Center, Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland.
58. Program in Genetics and Genome Biology, Hospital for Sick Children and University of Toronto, Toronto, Ontario, Canada.
59. Gen-info Ltd, Zagreb, Croatia. .
60. University of Split School of Medicine, Soltanska 2, Split, Croatia.
61. Ulverscroft Vision Research Group, University College London, London, UK.
62. Research Centre of Applied and Preventive Cardiovascular Medicine, University of Turku, Turku, Finland.
63. Department of Clinical Physiology and Nuclear Medicine, Turku University Hospital, Turku, Finland.
64. Myopia Research Group, Singapore Eye Research Institute, Singapore National Eye Centre, Singapore.
65. Saw Swee Hock School of Public Health, National University Health Systems, National University of Singapore, Singapore.
66. Department of Ophthalmology, University of Pennsylvania, Philadelphia, Pennsylvania, USA.
67. Department of Genetics, Genomics and Informatics, University of Tennessee Health Sciences Center, Memphis, Tenessee.
68. Department of Ophthalmology, University of Helsinki and Helsinki University Hospital, Helsinki, Finland.
69. Beijing Tongren Eye Center, Beijing Key Laboratory of Intraocular Tumor Diagnosis and Treatment, Beijing Ophthalmology & Visual Sciences Key Lab, Beijing Tongren Hospital, Capital Medical University, Beijing, China.
70. Department of Epidemiology and Medicine, Johns Hopkins Bloomberg School of Public Health, Baltimore, Maryland, USA.
71. Wilmer Eye Institute, Johns Hopkins Medical Institutions, Baltimore, Maryland, USA.
72. Department of Ophthalmology, Otsu Red Cross Hospital, Nagara, Japan. .
73. Centre for Myopia Research, School of Optometry, The Hong Kong Polytechnic University, Hong Kong, Hong Kong.
74. Department of Health Technology and Informatics, The Hong Kong Polytechnic University, Hong Kong, Hong Kong.
75. School of Ophthalmology and Optometry, Eye Hospital, Wenzhou Medical University, China.

3 23andMe

The following members of the 23andMe Research Team contributed to this study:

Michelle Agee

Babak Alipanahi

Adam Auton

Robert K. Bell

Katarzyna Bryc

Sarah L. Elson

Pierre Fontanillas

Nicholas A. Furlotte

David A. Hinds

Karen E. Huber

Aaron Kleinman

Nadia K. Litterman

Jennifer C. McCreight

Matthew H. McIntyre

Joanna L. Mountain

Elizabeth S. Noblin

Carrie A.M. Northover

Steven J. Pitts

J. Fah Sathirapongsasuti

Olga V. Sazonova

Janie F. Shelton

Suyash Shringarpure

Chao Tian

Joyce Y. Tung

Vladimir Vacic

Catherine H. Wilson.

4 Cohort Description

4.1 UK Biobank

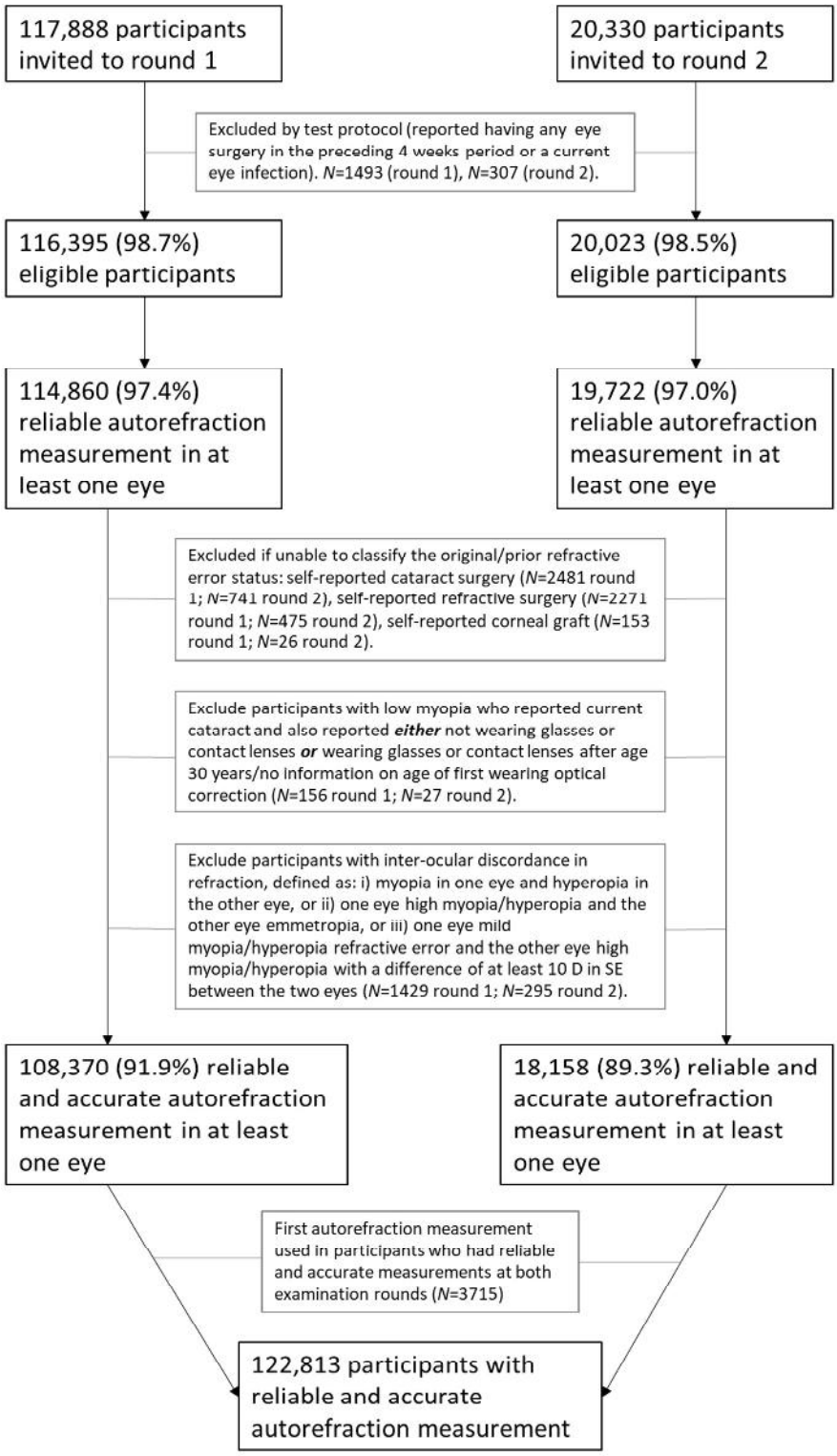
The UK Biobank is a very large multisite cohort study established by the Medical Research Council, Department of Health, Wellcome Trust medical charity, Scottish Government and Northwest Regional Development Agency. A baseline questionnaire, measurements, and biological samples were undertaken in 22 assessment centers across the UK between 2006 and 2010.

4.1.1 Phenotyping

4.1.1.1 Spherical equivalent (UK Biobank – 1)

Ophthalmic assessment was not part of the original baseline assessment and was introduced as an enhancement in 2009 for 6 assessment centers which are spread across the UK (Liverpool and Sheffield in North England, Birmingham in the Midlands, Swansea in Wales, and Croydon and Hounslow in Greater London). Participants completed a touch-screen self-administered questionnaire. The response options for ethnicity included White (English/Irish or other white background), Asian or British Asian (Indian/Pakistani/Bangladeshi or other Asian background), Black or Black British (Caribbean, African, or other black background), Chinese, mixed (White and Black Caribbean or African, White and Asian, or other mixed background), or other, non-defined, ethnic group.

Refractive error (RE) was measured by non-cycloplegic autorefraction in both eyes using the Tomey RC 5000 Auto Refkeratometer (Tomey Corp., Nagoya, Japan). The right eye was measured first and up to 10 measurements were taken per eye. The most representative result was automatically recorded. To ensure reliable and accurate RE data, we excluded participants based on previously published criteria[1] (Supplementary Note Figure 1). Spherical equivalent was calculated as spherical refractive error (UK Biobank codes 5084 and 5085) plus half the cylindrical error (UK Biobank 5086 and 5087) for each eye. If reliable data were only available for one eye, the RE of that eye was considered as the participant's RE. If reliable data were available for both eyes, we calculated the mean of right and left RE as the participant's RE.



Supplementary Note Figure 1. Flowchart explaining the exclusion of UK Biobank participants, as described elsewhere[1], for whom spherical equivalent measurements were available, based on existing clinical data.

A summary of the basic demographic characteristics of the subset of the UK Biobank for which the spherical equivalent was directly measured and which was used for our analysis is given below (Supplementary Note Table 1).

		Participants	Spherical Equivalent (diopters)	Age (years)
		N (%)	Mean ± SD	Mean ± SD
All		102,117	-0.28D (2.74)	57.26 (7.86)
Sex	**Male**	47,774	-0.27D (2.65)	57.59 (7.92)
	Female	54,343	-0.29D (2.82)	56.97(7.79)

Supplementary Note Table 1**.** Characteristics of UK Biobank participants for whom spherical equivalent measurements were available and included in the quantitative spherical equivalent analysis.
Abbreviations: N, number; SD, standard deviation; Age, age at first documented spherical equivalent assessment

4.1.1.2 Inferring myopia case-control status from self-reported age of spectacle wear (UKB2 sample)

Direct measurements of RE were only available for just 22.7% of the entire UK Biobank sample. However refractive error is strongly correlated with several clinical parameters and demographic factors[2], which can be used to predict refractive error and myopia. Some of that indirect information was present for significant numbers among the UK Biobank participants.

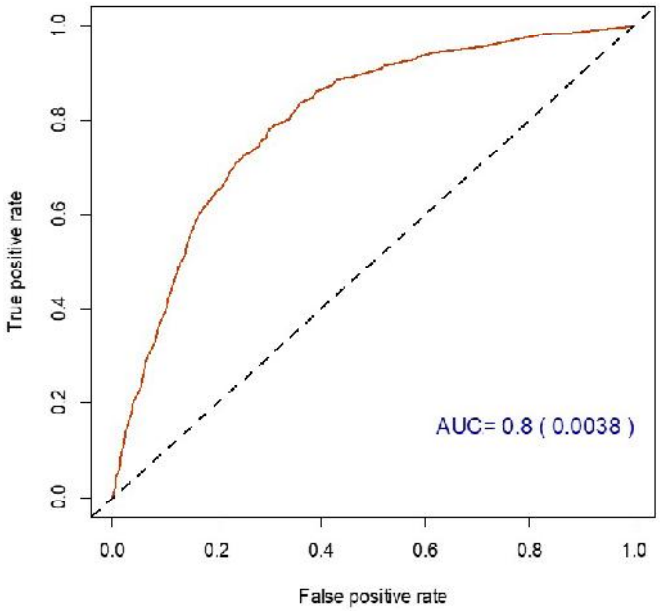
For example, age when the first lens correction is prescribed is strongly correlated to spherical equivalent both at a genetic[3] and phenotypic level[4]. It has used before as a proxy for RE previously[5] with reasonably low levels of genetic effect heterogeneity[6].A total of 87% of the participants, responded to the question on whether their vision needed correction when they were asked "Do you wear glasses or contact lenses to correct your vision?" (potential answers: Yes/No/Prefer not to answer) in a touch-screen self-administered questionnaire. If a participant answered "Yes" to this question, they were further asked "What age did you first start to wear glasses or contact lenses?", to which 67% of the participants responded. In addition, spherical equivalent and myopia affection status is highly correlated with age, sex, birth year[4,7] , all of which were available for UK Biobank participants.

We therefore aimed at harnessing the demographic and clinical information to obtain an estimate about the individual's likely myopia status. This general approach has been used successfully before[6], and to better classify the non-refracted subjects into myopia cases and non-myopia controls.

We proceeded in three steps: 1) training a Support Vector Machine model in 80% randomly selected UK Biobank participants of European descent for whom direct spherical equivalent and refractive error

status were available, 2) validate the prediction in the remaining 20% of UK Biobank participants of European descent for whom direct spherical equivalent and refractive error status were available, 3) make the SVM predictions in the remaining individuals with no direct spherical error measurements available using the model developed for the training data.

We initially fine-tuned the prediction model, in order to optimize the γ (gamma) and "cost" parameters in the 80% training data sample. The performance of the model was subsequently tested in the remaining 20% of the UK Biobank participants with spherical equivalent measurements. Receiver Operating Characteristic (ROC) curves were drawn and the area under the curve (AUC) calculated (Supplementary Note Figure 2). For the optimal SVM model, an Area Under the Curve (AUC) of 0.8 was obtained, providing strong evidence that myopia status could be inferred from 'age of first spectacle wear' with sufficient reliability to serve as a proxy for myopia status in a GWAS analysis. When we compared predicted vs observed myopia case-control status in the 20% validation subset of the UK Biobank for which spherical equivalent were available (cases defined as < -0.75 Diopters and controls > -0.5Diopters), the inferred myopia status was a strong predictor for actual (observed) myopia case-control status (OR=16.79, 95%CI 15.99-17.62).



Supplementary Note Figure 2. ROC curve and AUCs for the SVM model trained in 80% of fully phenotyped (i.e. spherical equivalent available) UK Biobank participants. For both ROC and AUC, the remaining 20% (i.e. not part of the initial training set) of the UK Biobank were used. Cases defined as < -0.75 Diopters and controls > -0.5Diopters. The standard error for the AUC is shown within brackets.

The optimal SVM model was applied to the UK Biobank sample with known 'age of first spectacle wear' but for whom no spherical equivalent measurements were available. That is, the GWAS for SVM-inferred myopia case-control status did not include participants with known (measured) refractive error. This inferred phenotypic status was used for a GWAS analysis, for which we followed the previously

described workflow (sections 4.1.3). In further support of SVM-inferred myopia case-control status as a valid proxy measure, the genetic effect sizes for variants assessed in the SVM-inferred phenotype were found to be highly correlated to those from the GWAS analyses for autorefractor-measured refractive error (Supplementary Note Table 2, Supplementary File 1**Error! Reference source not found.**).

	UKB1[1]	GERA[1]	23andMe[2,3]	UKB2[2,4]
Beta - UKB1	1.000	0.926	0.953	0.935
Beta - GERA	0.926	1.000	0.927	0.901
23andMe - log(OR)	0.953	0.927	1.000	0.941
UKB-2 log(OR)	0.935	0.901	0.941	1.000

Supplementary Note Table 2. Correlation of effects sizes between the UKB-2 subset (for which spherical equivalent was not available and myopia status was imputed using the SVM) with other cohorts in which refraction was directly measured or self-reported. "Beta" linear regression coefficients; log(OR) the logarithm of the logistic regression Odds Ratios; UKB-1 denotes the first subset of the UK Biobank participants (spherical equivalent available), UKB -2 the second subset (myopia case-control status inferred). For a description of the 23andMe and GERA cohorts, please refer to subsequent sections for further cohort descriptions.

The basic characteristics of the participants that were selected for the second UK Biobank subset (UKB-2) of our meta-GWAS are shown in the Supplementary Note Table 3.

		inferred cases/controls	**Age first spectacle wear (years)**	**Age (years)**
		N cases (%)	Mean ± SD	Mean ± SD
All		108,956/70,941 (60.56)	26.1 ± 13.58	56.61 ± 7.79
Sex	**Male**	45,994/30,388 (60.25)	26.38 ± 13.48	56.78 ± 7.88
	Female	62,962 / 40,553 (60.82)	25.74 ± 13.65	56.48 ± 7.73

Supplementary Note Table 3**.** Characteristics of UK Biobank participants included in the inferred myopia qualitative analysis (UKB-2). Abbreviations: N, number; SD, standard deviation; Age, age at first documented spherical equivalent assessment.

4.1.2 Genotyping

DNA extraction, genotyping and imputation of UK Biobank participants has been reported elsewhere[8]. DNA extraction begun on buffy coat samples. DNA was extracted from 850 µl buffy coat (recovered from 9 ml of whole blood) on customized TECAN Freedom EVO® 200 platform[9]. The samples were then processed in the approximate order received to produce genotype data. Genotyping was done using two arrays. The first array was the Affymetrix Axiom® platform with a custom-designed array described in the UK Biobank Axiom® Array Content Summary[10]. Processing was done using a LIMS system to track instrumentation, Axiom consumables arrays and reagents and operators. The process is described

elsewhere[11]. Details on genotyping procedure and quality control can be found elsewhere[12]. The second array is what was the UK BiLEVE, described elsewhere[13].

Phasing on the autosomes was carried out using a modified version of the SHAPEIT2[14] program modified to allow for very large sample sizes. This new method (which we refer to as SHAPEIT3) modifies the SHAPEIT2 surrogate family approach to remove a quadratic complexity component of the algorithm[15]. In small sample sizes of a few thousand samples, this part of the algorithm, which involves calculating Hamming distances between current haplotypes estimates, contributes only a relatively small part to the computational cost. As sample sizes increase over 10,000 samples then this component becomes significant. The new algorithm uses a divisive clustering algorithm to identify clusters of haplotypes, and then calculates Hamming distances only between pairs of haplotypes within each cluster. Only haplotypes within each cluster are used as candidates for the surrogate family copying states in the HMM model.

A total of 806,466 directly genotyped DNA sequence variants were available after variant quality control. The UK Biobank team then performed imputation from a combined Haplotype Reference Consortium (HRC) and UK10K reference panel; phasing was performed using SHAPEIT3 and imputation was carried out via the IMPUTE4 program[16]. Only HRC-imputed variants were used for the purpose our analyses of the UK Biobank participants. The variant-level quality control exclusion metrics applied to imputed data for GWAS included the following: call rate < 95%, Hardy–Weinberg equilibrium $P < 1 \times 10^{-6}$, posterior call probability < 0.9, imputation quality < 0.4, and MAF < 0.005. The Y chromosome and mitochondrial genetic data were excluded from this analysis. In total, 10,263,360 imputed DNA sequence variants were included in our analysis.

For sample quality control, we removed individuals of non-European ancestry and participants with relatedness corresponding to third-degree relatives or closer, and an additional 480 samples with an excess of missing genotype calls or more heterozygosity than expected were excluded. In total, genotypes were available for 102,117 participants of European ancestry with spherical equivalent data.

4.1.3 Association analyses

The basic model tested was the average of spherical equivalent measured in the left and right eye as an outcome of a regression model whose predictor is the allele dosage at a given polymorphic locus, adjusted for the effect of relevant covariables (see table below). The empirical association between spherical equivalent and other covariables is shown in Supplementary Note Table 4.

Effect size estimates (Beta) and P-values are from the multivariate regression model. Since demographic factors and principal components had a small yet real effect over Spherical Equivalent, the above variables were included in the model.
Therefore, models of mixed linear regressions, as described before[17], where the spherical equivalent was the outcome, the allele dosage the predictor, adjusted for age, sex and the first 10 principal components. Since there was, there is evidence of cryptic relatedness among the UK Biobank participants, a linear mixed model that controls for population structure was used as implemented in the Bolt-LMM software [18].

Variable	Beta	SE	Z	P-value
age	0.05247	0.001122	46.773	<E-308
sex	-0.08013	0.017988	-4.455	8.42E-06
PC1	0.000297	0.000125	2.367	0.017935
PC2	0.001096	0.000272	4.032	5.54E-05
PC3	0.001369	0.000516	2.655	0.007935
PC4	0.005147	0.000799	6.439	1.21E-10
PC5	0.003194	0.001253	2.55	0.010786
PC6	0.004701	0.001314	3.577	0.000348
PC7	0.008073	0.001614	5.002	5.68E-07
PC8	0.000269	0.001605	0.168	0.866871
PC9	0.008165	0.002385	3.424	0.000618
PC10	-0.00039	0.001885	-0.207	0.836117

Online Note Table 4. The association between spherical equivalent, age, sex and the first 10 Principal Components.

For the second UK Biobank subset, for which no spherical equivalent information was available, the mixed linear model was built with the predicted myopia status as outcome and using the same covariates as for the previously described linear regression analysis on spherical equivalent (paragraph 4.1.3). Odds Ratios were obtained from the beta regression coefficient using the equation:

$$\ln(\mathrm{OR}) = \frac{\beta}{\mu(1-\mu)}$$

where μ is the fraction of the cases in the sample (μ=0.606). Although the case-control analysis was quite balanced, we opted to remove genotypes with MAF <0.01 and MAC< 400 recommended elsewhere[19] (which in our samples, most often would be correspond to MAF < 0.001).

4.2 23andMe

4.2.1 Phenotyping

The subjects were all volunteers from the 23andMe (Mountain View, CA, USA) personal genomics company. All participants included in the analyses provided informed consent and answered surveys online according to the approved 23andMe human subjects protocol, which was reviewed and approved by Ethical & Independent Review Services, a private institutional review board (http://www.eandireview.com). The participants were identified as myopia cases if they responded positively to any of the following questions:

1. "Have you ever been diagnosed by a doctor with nearsightedness (near objects are clear, far objects are blurry)?"
2. "Are you nearsighted (near objects are clear, far objects are blurry)?"
3. "What vision problems do you have? Please check all that apply." - Nearsightedness (near objects are clear, far objects are blurry.
4. "Prior to your LASIK eye surgery, what vision problems did you have? Please check all that apply." - Nearsightedness (near objects are clear, far objects are blurry.

Controls were defined as having said "No" or not checking nearsightedness to at least one of the questions above. Subjects who gave discordant answers were removed.

4.2.2 Genotyping

DNA extraction and genotyping were performed on saliva samples by CLIA-certified and CAP-accredited clinical laboratories of Laboratory Corporation of America. Samples were genotyped on one of four genotyping platforms. The V1 and V2 platforms were variants of the Illumina HumanHap550+ BeadChip, including about 25,000 custom SNPs selected by 23andMe, with a total of about 560,000 SNPs. The V3 platform was based on the Illumina OmniExpress+ BeadChip., with custom content to improve the overlap with our V2 array, with a total of about 950,000 SNPs. The V4 platform in current use is a fully custom array, including a lower redundancy subset of V2 and V3 SNPs, with additional coverage of lower-frequency coding variation, and about 570,000 SNPs. Samples that failed to reach 98.5% call rate were re-analyzed. For the GWAS only participants who have >97% European ancestry, as determined through an analysis of local ancestry, were included. For the purposes of ethnic categorization, an algorithm first partitioned phased genomic data into short windows of about 100 SNPs and used a support vector machine (SVM) to classify individual haplotypes into one of 31 reference populations. The SVM classifications then fed into a hidden Markov model (HMM) that accounts for switch errors and incorrect assignments and gives probabilities for each reference population in each window. The reference population data are derived from public datasets (the Human Genome Diversity Project, HapMap, and 1000 Genomes), as well as 23andMe customers who have reported having four grandparents from the same country. A maximal set of unrelated individuals was chosen for each analysis using a segmental identity-by-descent (IBD) estimation algorithm[20]. Individuals were defined as related if they shared more than 700 cM IBD, including regions where the two individuals share either one or both genomic segments identical-by-descent. This level of relatedness corresponds approximately to the minimal expected sharing between first cousins in an outbred population.

Participant genotype data were imputed against the September 2013 release of 1000 Genomes Phase1 reference haplotypes, phased with ShapeIt2[21]. We phased and imputed data for each genotyping platform separately. We phased using an internally developed phasing tool which implements the Beagle haplotype graph-based phasing algorithm[22].

SNPs with Hardy-Weinberg equilibrium $P<10^{-20}$, call rate < 95%, or with large allele frequency discrepancies compared to European 1000 Genomes reference data were excluded from imputation. Imputation was done against all-ethnicity 1000 Genomes haplotypes (excluding monomorphic and singleton sites) using Minimac[23]. For the X chromosome, separate haplotype graphs were built for the non-pseudoautosomal region and each pseudoautosomal region, and these regions were phased separately. Males and females were imputed together using Minimac2[23], as with the autosomes, treating males as homozygous pseudo-diploids for the non-pseudoautosomal region.

HLA allele dosages were imputed from SNP genotype data using HIBAG[24]. We imputed alleles for HLA-A, B, C, DPB1, DQA1, DQB1, and DRB1 loci at four-digit resolution. To test associations between HLA allele dosages and phenotypes, we performed logistic or linear regression using the same set of covariates used in the SNP-based GWAS for that phenotype. We performed separate association tests for each imputed allele.

4.2.3 Association analyses

Association test results were computed by linear regression assuming additive allelic effects. For tests imputed dosages rather than best-guess genotypes were used. Covariates for age, gender, the first ten principal components to account for residual population structure were also included into the model. Results for the X chromosome are computed similarly, with male genotypes coded as if they were homozygous diploid for the observed allele.

4.3 GERA

The Genetic Epidemiology Research in Adult Health and Aging (GERA) cohort is part of the Kaiser Permanente Research Program on Genes, Environment, and Health (RPGEH) and has been described in detail elsewhere[25,26]. The GERA cohort comprises 110,266 adult men and women who are consented participants in the RPGEH, an unselected cohort of adult participants who are members of Kaiser Permanente Northern California (KPNC), an integrated health care delivery system, with ongoing longitudinal records from vision examinations. For this analysis, 34,998 adults (25 years and older), who self-reported as non-Hispanic white, and who had at least one assessment of spherical equivalent obtained between 2008 and 2014 were included (Supplementary Note Table 5). All study procedures were approved by the Institutional Review Board of the Kaiser Foundation Research Institute.

		Participants	**Spherical Equivalent (diopters)**	**Age (years)**
		N (%)	Mean ± SD	Mean ± SD
All		34,998 (100)	-0.35 ± 2.56	66.54 ± 11.55
Sex	**Male**	14,431 (41.23)	-0.32 ± 2.46	68.84 ± 10.70
	Female	20,567 (58.77)	-0.38 ± 2.64	64.93 ± 11.84

Supplementary Note Table 5**.** Characteristics of GERA non-Hispanic white subjects included in the GWAS of spherical equivalent by sex. Abbreviations: N, number; SD, standard deviation; Age, age at first documented spherical equivalent assessment

4.3.1 Phenotyping

All participants underwent vision examinations, and most subjects had multiple measures for both eyes. Spherical equivalent was assessed as the sphere + cylinder/2. For this analysis, spherical equivalent was selected from the first documented assessment, and the mean of both eyes was used. As previously described[27], individuals with histories of cataract surgery (in either eye), refractive surgery, keratitis, or corneal diseases were excluded.

4.3.2 Genotyping

DNA samples from GERA individuals were extracted from Oragene kits (DNA Genotek Inc., Ottawa, ON, Canada) at KPNC and genotyped at the Genomics Core Facility of the University of California, San Francisco (UCSF). DNA samples were genotyped at over 665,000 single nucleotide polymorphisms (SNPs) on Affymetrix Axiom arrays (Affymetrix, Santa Clara, CA, USA)[28,29]. SNPs with initial genotyping call rate ≥97%, allele frequency difference ≤0.15 between males and females for autosomal markers, and genotype concordance rate >0.75 across duplicate samples were included[26]. Around 94% of samples and more than 98% of genetic markers assayed passed quality control (QC) procedures. In addition to those QC criteria,

SNPs with genotype call rates <90% were removed, as well as SNPs with a minor allele frequency (MAF) < 1%.
Following genotyping QC, we conducted statistical imputation of additional genetic variants. Following the pre-phasing of genotypes with Shape-IT v2.r72719[30], variants were imputed from the cosmopolitan 1000 Genomes Project reference panel (phase I integrated release; http://1000genomes.org) using IMPUTE2 v2.3.0.[31-33] As a QC metric, we used the info r^2 from IMPUTE2, which is an estimate of the correlation of the imputed genotype to the true genotype[34]. Variants with an imputation $r^2 < 0.3$ were excluded, and we restricted to SNPs that had a minor allele count (MAC) ≥ 20.

4.3.3 Association analyses

A linear regression of each individual's spherical equivalent was performed with the following covariates: age at first documented spherical equivalent assessment, sex, and genetic principal components. A linear regression of the residuals on each SNP was then performed using PLINK[35] v1.9 (www.cog-genomics.org/plink/1.9/) to assess genetic associations. Data from each SNP were modeled using additive dosages to account for the uncertainty of imputation[36]. Eigenstrat[37] v4.2 was used to calculate the PCs[25]. The top 10 ancestry PCs were included as covariates, as well as the percentage of Ashkenazi ancestry to adjust for genetic ancestry, as described previously[25].

4.4 Cream Consortium

4.4.1 Phenotyping

All participants included in this analysis from CREAM were 25 years of age or older. RE was represented by measurements of refraction and spherical equivalent (SphE = spherical refractive error +1/2 cylinder refractive error) was the outcome variable for CREAM. Participants with conditions that might alter refraction, such as cataract surgery, laser refractive procedures, retinal detachment surgery, keratoconus, or ocular or systemic syndromes were excluded from the analyses. Recruitment and ascertainment strategies varied by study and were previously published elsewhere[6].

4.4.2 Genotyping

The genotyping process has been described elsewhere[6]. Samples were genotyped on different platforms, and study-specific QC measures of the genotyped variants were implemented before association analysis. Genotypes were imputed with the appropriate ancestry-matched reference panel for all cohorts from the 1000 Genomes Project (Phase I version 3, March 2012 release) with either minimac[23] or IMPUTE[16]. The metrics for preimputation QC varied among studies, but genotype call-rate thresholds were set at a high level (≥0.95). These metrics were similar to those described in a previous GWAS analyses[38]; detailed information for each cohort is described elsewhere[6].

4.4.3 Association analyses

To prevent overlap of samples, cohorts from the United Kingdom (1985BBC, ALSPAC-Mothers, EPIC-Norfolk, ORCADES and Twins UK) were excluded from the GWAS meta-analysis. Association analyses were performed following the workflow elsewhere[6]: All samples analyzed were of European descent, for each CREAM cohort, a single-marker analysis for the phenotype of Spherical equivalent (in diopters) was carried out with linear regression with adjustment for age, sex and up to the first five principal components. For all non-family-based cohorts, one of each pair of relatives was removed (after detection through either GCTA or identity by sequence (IBS)/identity by descent (IBD) analysis). In

family-based cohorts, a score test-based association was used to adjust for within-family relatedness. We used an additive SNP allelic-effect model.

4.5 EPIC

The European Prospective Investigation into Cancer (EPIC) study is a pan-European prospective cohort study designed to investigate the etiology of major chronic diseases[39]. EPIC-Norfolk , one of the UK arms of EPIC, recruited and examined 25,639 participants between 1993 and 1997 for the baseline examination[40]. Recruitment was via general practices in the city of Norwich and the surrounding small towns and rural areas, and methods have been described in detail previously[41]. Since virtually all residents in the UK are registered with a general practitioner through the National Health Service, general practice lists serve as population registers. Ophthalmic assessment formed part of the third health examination and this has been termed the EPIC-Norfolk Eye Study[42]. In total, 8,623 participants were seen for the Eye Study, between 2004 and 2011. The EPIC-Norfolk Eye Study was carried out following the principles of the Declaration of Helsinki and the Research Governance Framework for Health and Social Care. The study was approved by the Norfolk Local Research Ethics Committee (05/Q0101/191) and East Norfolk & Waveney NHS Research Governance Committee (2005EC07L). All participants gave written, informed consent.

4.5.1 Phenotyping

Refractive error was measured in both eyes using a Humphrey Auto-Refractor 500 (Humphrey Instruments, San Leandro, California, USA). Spherical equivalent was calculated as spherical refractive error plus half the cylindrical error for each eye.
Some basic demographic and clinical information about the samples used for the validation analyses is given below (Supplementary Note Table 6).

		Participants	**Spherical Equivalent (diopters)**	**Age (years)**
		N (%)	Mean ± SD	Mean ± SD
All		7,117 (100)	+0.16 ± 2.25	68.80 ± 8.18
Sex	**Male**	3,253 (45.71)	+0.15 ± 2.23	69.60 ± 8.20
	Female	3,864 (54.29)	+0.18 ± 2.27	68.13 ± 8.11

Supplementary Note Table 6**.** Characteristics of the participants in the EPIC-Norfolk cohort, included in the heritability and risk prediction analyses. Abbreviations: N, number; SD, standard deviation; Age, age at first documented spherical equivalent assessment

4.5.2 Genotyping and imputation

Genotypes obtained using the Affymetrix UK Biobank Axiom Array on 7,117 subjects contributed to the current study were excluded if they had low call rates, poor clustering, batch effects across genotyping plates and/or Hardy-Weinberg equilibrium $P < 10^{-7}$. Samples were excluded on grounds of poor genotyping across all SNPs, sex discordance , exces or low heterozygosity and unexplainable identity-by-descent values. Third-degree relatives or closer participants were also removed. Data were pre-phased using SHAPEIT[14] version 2 and imputed to the Phase 3 build of the 1000 Genomes project[43] (October 2014) using IMPUTE[16] version 2.3.2.

4.5.3 Association analysis

We examined the relationship between allele dosage and mean spherical equivalent using linear regression adjusted for age, sex and the first 5 principal components. Analyses were carried out using SNPTEST version 2.5.1.

5 STATISTICAL ANALYSES

5.1 Meta-analyses

For all meta-analyses we applied a Z-score method, weighted by the effective population sample size, as implemented in METAL[44]. No genomic control adjustment was applied during the meta-analysis.

5.2 Conditional analyses

The conditional and joint analysis on summary data (COJO) [45] as implemented in the GCTA program [46] was used to identify independent effects within associated loci as well as the calculation of the phenotypic variance explained[47] by all polymorphisms associated with the trait after the conditional analyses. Default parameters were used for the analysis. The LD estimates were derived from a randomly selected sample of 10,000 unrelated subjects the UK Biobank cohort.

5.3 Multiple testing correction

Two methods of correcting for multiple testing were used. The first was a classic Bonferroni correction, in which the threshold of significance (0.05) was divided by the number of tests (n):

$$\alpha = \frac{0.05}{n}$$

Given the large number of loci for which replication was needed, we additionally calculated the False Discovery Rates, using the Benjamini-Hochberg method[48].

5.4 Genomic inflation

To assess the potential inflation of association probabilities, genomic inflation factors[49] were calculated and Q-Q plots were drawn using the package 'gap' in R (https://cran.r-project.org/).

5.5 LDscore regression-based methods

5.5.1 Polygenicity vs. inflation

To distinguish between the effect of polygenicity and those arising from sample stratification or uncontrolled population admixture, we followed previously suggested approaches[50] to calculate the LD score regression intercepts using the program LD Score (https://github.com/bulik/ldsc).

5.5.2 Calculation of genetic correlation

Bivariate genetic correlations between refractive error and other complex traits whose summary statistics are publicly available were assessed following previously described methodologies[51], using the program LD Score (https://github.com/bulik/ldsc).

5.6 Associated SNPs and gene annotations

Polymorphisms associated at a GWAS level (P<5x10^{-08}) were clustered within an "associated genomic region", defined as a contiguous genomic region where GWAS-significant markers were within 1 million base pairs from each other, as suggested elsewhere[52]. Significant polymorphisms were annotated with the gene inside whose transcript-coding region they are located, or alternatively, if located between two genes, with the gene nearest to it. The associated genomic regions were collectively annotated with the gene overlapping, or nearest the most significantly associated variant within that region. In addition, the polymorphic sites were functionally annotated using SNPnexus[53].

5.6.1 OMIM

The Online Mendelian Inheritance In Man (OMIM) is a continuously curated catalog of human genes and phenotypic changes their polymorphic forms cause in humans[54]. This catalogue contains a still partial, but highly reliable list of gene-phenotype pairs and was used retrieve data that could inform about the functionality of specific genes with particular focus on phenotypic expressions of extremely penetrant mutations.

5.6.2 The GWAS Catalog.

Previous GWAS association of SNPs or genes with other phenotypic traits was conducted through queries of the GWAS Catalog[55]. Results were downloaded from the official site hosted at the European Bioinformatics Institute: https://www.ebi.ac.uk/gwas/downloads.

5.7 Graphical illustration of association

LocusZoom[56] was used to generate plot that visualize regional association and its genomic context. Data from the European participants in the 1000 Genome Project, November 2014 was used, and the graphs were generated using the online LocusZoomserver (http://locuszoom.org/).

5.8 Mendelian randomization

The R (https://cran.r-project.org) package MendelianRandomization v3.4.4 was used for Mendelian randomization analyses.

5.9 Gene expression, GTEx and other transcription data

We obtained data on tissue expression from several sources for genes that map within RE associated loci defined as described before (section **Error! Reference source not found.**). Information about the expression of the genes of interest in systemic (i.e. non-ocular) tissues was obtained from the GTEx Portal for GTEx release v7 (https://gtexportal.org/home/datasets). RNA sequencing data was obtained

for both fetal and adult corneal, trabecular meshwork and ciliary body, as described elsewhere[57], which we downloaded from the authors' supplementary information. In addition, we extracted data from the subset of subjects with presumed healthy adult retinas (AMD=1), described elsewhere[58] that obtained from the GTEx Portal (https://gtexportal.org/home/datasets).

Transcription data was processed using different platforms and were available in different units (Transcripts per Million bases, TPM, for the retina and GTEx tissues, and Fragments per Kilobase, FPKM for the other tissues). For purposes of comparing expression across different tissues for which different methodologies may have been used, expression levels for all tissues were rank-transformed. Hierarchical clustering was used to help visualize similarities and differences of patterns of transcript expression across different tissues ('hclust' package in R).

5.10 LD score regression applied to specifically expressed genes (LDSC-SEG)

Disease-relevant tissues and cell types were identified by analyzing gene expression data together with summary statistics from the meta-analysis of refractive error in all five cohorts, as described elsewhere[59]. Briefly, genes were ranked based on the t-statistic of their expression in each tisue and the 10% most expressed genes for each tissue were considered "specifically expressed genes". A stratified LD score regression was applied to the meta-analysis summary statistics to evaluate the contribution of the focal genome annotation to trait heritability.

5.11 SMR

SMR (Summary data–based Mendelian randomization) assesses the relationship between genetic variant, intermediate variables such as gene expression levels or methylation levels as mediating traits, to test causality on a specific phenotype[60].

5.11.1 Test description

The SMR package helps perform two tests. The first is an SMR test, which correlates GWAS effects with eQTL or methylation effects (or any other intermediate trait)[60]. This test suggests causation, although it is unable to fully differentiate between it and pleiotropy. The second test is that of Heterogeneity in Dependent Instrument (HEIDI). This test against the null hypothesis that changes in both eQTL (or other intermediary traits) and the phenotype of interest are caused by one single SNP, which is therefore considered as the candidate for the putative causal effect.

5.11.2 Datasets for the SMR analyses: eQTL, *cis*-mQTL

To perform the above-mentioned tests of causation/pleiotropy, we used three different datasets of association between genetic variants and intermediate traits. The first was the summary statistics of eQTL associations in the untransformed peripheral blood samples of 5,311 subjects[61]. There were two advantages in using these data: 1) this was the largest eQTL dataset available and 2) the use of a highly heterogeneous tissue such as peripheral blood would be more likely than any other single more homogeneous tissue to overcome any heterogeneity of eQTL effects with eye and retinal tissues that were unavailable at the time of the analysis and manuscript writing.

Assuming that tissues relevant to the development of refractive error are similar to the brain, we also used two datasets, one with eQTL effects and the other with results of a *cis*- methylation analysis (*cis*-mQTL), both in brain tissues[62].

5.12 Gene-set enrichment

To identify pathways or other gene sets that were over-represented among our results, we used a Gene-Set Enrichment Analysis (GSEA) as implemented in the Meta-Analysis Gene Set Enrichment of Variant (MAGENTA) software[63]. This program assigns scores to each gene based on the strength of association with refractive error, adjusting for potential confounders such as gene length and linkage disequilibrium. Enrichment for any gene set was assessed within genes above the cut-off of the highest 75th centile of significant gene scores. For the current study, the most recent versions of Gene Ontology (GO), Panther, KGG, Biocarta and MSigDB databases were used. We also carried out a similar enrichment analysis for the presence of transcription factor binding sites. A permutational procedure and false-discovery rates were used to calculate significance of enrichment and control for multiple testing.

5.12.1 GSEA definitions

For the enrichment analyses we used updated versions of the GSEA gene sets as described before[64]. We used the versions from September 2017 which were downloaded from:
http://software.broadinstitute.org/gsea/login.jsp

5.13 Analyses of signals of natural selection

Results of three statistical tests for natural selection were imported from the 1000 Genomes Selection Browser[65]. We downloaded and reported results from several tests such as iHS[66] and a cross-population comparison, XP-EHH, based on extended haplotype homozygosity test (average and maximum CEU, CEU vs YRI) [67] and the Tajima's D test. The absolute test scores and the rank scores (-log10 of the centile of the absolute test score across the genome) were reported.

5.14 Estimation of effect size distributions for spherical equivalent

We used a maximum-likelihood model to estimate the distribution of effect sizes, based on summary statistics of observations and linkage disequilibrium patterns to predict the likely number of SNPs that explain spherical equivalent heritability as well as explore the relationship between future sample sizes and the number of SNPs identified and variance or heritability explained as described elsewhere[68] and implemented in the GENESIS R package (https://github.com/yandorazhang/GENESIS).

6 FULL ACKNOWLEDGEMENTS

6.1 UK Biobank

UK Biobank was established by the Wellcome Trust medical charity, Medical Research Council, Department of Health, Scottish Government, and Northwest Regional Development Agency. It also had funding from the Welsh Assembly Government, British Heart Foundation, and Diabetes UK.

Development of the eye and vison dataset was led by Prof Sir Peng T Khaw and Prof Paul Foster, funded by a grant from The National Institute for Health Research (NIHR) to Moorfields Eye Hospital NHS Foundation Trust and UCL Institute of Ophthalmology for a biomedical research centre.

6.2 23andMe

23andMe would like to thank the research participants and employees of 23andMe for making this work possible.

6.3 GERA

GERA are grateful to the Kaiser Permanente Northern California members who have generously agreed to participate in the Kaiser Permanente Research Program on Genes, Environment, and Health. Support for participant enrollment, survey completion, and biospecimen collection for the RPGEH was provided by the Robert Wood Johnson Foundation, the Wayne and Gladys Valley Foundation, the Ellison Medical Foundation, and Kaiser Permanente Community Benefit Programs. Genotyping of the GERA cohort was funded by a grant from the National Institute on Aging, National Institute of Mental Health, and National Institute of Health Common Fund (RC2 AG036607). Data analyses were facilitated by National Eye Institute (NEI) grant R01 EY027004 (E.J.), National Institute of Diabetes and Digestive and Kidney Diseases grant R01 DK116738 (E.J.).

6.4 Consortium for Refractive Error and Myopia (CREAM)

CREAM investigators are gratefully thank all study participants, their relatives and the staff at the recruitment centers for their invaluable contributions. Funding for this particular GWAS mega-analysis was provided by the European Research Council (ERC) under the European Union's Horizon 2020 Research and Innovation Program (grant 648268), the Netherlands Organisation for Scientific Research (NWO, grant 91815655) and the National Eye Institute (grant R01EY020483). VJMV acknowledges funding from the Netherlands Organisation for Scientific Research (NWO, grant 91815655 & 91617076).

6.5 EPIC-Norfolk

EPIC-Norfolk infrastructure and core functions are supported by grants from the Medical Research Council (G1000143) and Cancer Research UK (C864/A14136). The clinic for the third health examination was funded by Research into Ageing (262). Genotyping was funded by the Medical Research Council (MC_PC_13048). We thank all staff from the MRC Epidemiology laboratory team for the preparation and quality control of DNA samples. Mr. Khawaja is supported by a Moorfields Eye Charity grant. Professor Foster has received additional support from the Richard Desmond Charitable Trust (via Fight for Sight) and the Department for Health through the award made by the National Institute for Health Research to

Moorfields Eye Hospital and the UCL Institute of Ophthalmology for a specialist Biomedical Research Centre for Ophthalmology.

6.6 King's College London authors

M.J. Simcoe is a recipient of a Fight for Sight PhD studentship. K. Patasova is a recipient of a Fight for Sight PhD studentship. P.G. Hysi the recipient of a FfS ECI fellowship. P.G. Hysi and C.J. Hammond acknowledge the TFC Frost Charitable Trust Support for the KCL Department of Ophthalmology. The statistical analyses were run in King's College London Rosalind HPC LINUX Clusters and cloud server. The UK Biobank data was accessed as part of the UK Biobank projects 669 and 17615.

6.7 UK Biobank Eye Consortium members

Data analyses were carried out using the RAVEN computing cluster, maintained by the ARCCA group. UK Biobank data was accessed as part of the UK Biobank Project 17615. Veronique Vitart is supported by core grant MC_UU_00007/10 from the UK Medical Research Council. The UK Biobank Eye and Vision Consortium has been supported by grants from NIHR (BRC3_026), Moorfields Eye Charity (ST 15 11 E), Fight for Sight (1507/1508), The Macular Society, The International Glaucoma Association (IGA, Ashford UK) and Alcon Research Institute.

6.8 University College London authors

Jugnoo Rahi is a National Institute for Health Research (NIHR) Senior Investigator, supported by the NIHR Biomedical Research Centres at Moorfields Eye Hospital/UCL Institute of Ophthalmology, and at the UCL Institute of Child Health/Great Ormond Street Hospital. Views expressed are those of the authors and not necessarily those of the NHS/NIHR. Phillippa Cumberland was funded by the Ulverscroft Foundation. Omar Mahroo is supported by Wellcome Trust grant 206619_Z_17_Z and the NIHR Biomedical Research Centre at Moorfields Eye Hospital and the UCL Institute of Ophthalmology. The UK Biobank data was accessed as part of the UK Biobank projects 669 and 17615.

References:

1. Cumberland, P.M. *et al.* Frequency and Distribution of Refractive Error in Adult Life: Methodology and Findings of the UK Biobank Study. *PLoS One* **10**, e0139780 (2015).
2. Williams, K.M. *et al.* Early life factors for myopia in the British Twins Early Development Study. *British Journal of Ophthalmology*, bjophthalmol-2018-312439 (2018).
3. Wojciechowski, R. & Hysi, P.G. Focusing in on the complex genetics of myopia. *PLoS Genet* **9**, e1003442 (2013).
4. Williams, K.M. *et al.* Age of myopia onset in a British population-based twin cohort. *Ophthalmic Physiol Opt* **33**, 339-45 (2013).
5. Kiefer, A.K. *et al.* Genome-wide analysis points to roles for extracellular matrix remodeling, the visual cycle, and neuronal development in myopia. *PLoS Genet* **9**, e1003299 (2013).
6. Tedja, M.S. *et al.* Genome-wide association meta-analysis highlights light-induced signaling as a driver for refractive error. *Nat Genet* **50**, 834-848 (2018).
7. Williams, K.M. *et al.* Increasing Prevalence of Myopia in Europe and the Impact of Education. *Ophthalmology* **122**, 1489-97 (2015).
8. Bycroft, C. *et al.* Genome-wide genetic data on~ 500,000 UK Biobank participants. *bioRxiv*, 166298 (2017).
9. http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/DNA-Extraction-at-UK-Biobank-October-2014.pdf.
10. http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014.pdf.
11. http://biobank.ctsu.ox.ac.uk/crystal/refer.cgi?id=155583.
12. https://biobank.ctsu.ox.ac.uk/crystal/docs/impute_ukb_v1.pdf.
13. Wain, L.V. *et al.* Novel insights into the genetics of smoking behaviour, lung function, and chronic obstructive pulmonary disease (UK BiLEVE): a genetic association study in UK Biobank. *Lancet Respir Med* **3**, 769-81 (2015).
14. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nat Methods* **9**, 179-81 (2011).
15. O'Connell, J. *et al.* Haplotype estimation for biobank-scale data sets. *Nat Genet* **48**, 817-20 (2016).
16. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet* **5**, e1000529 (2009).
17. Loh, P.R. *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* **47**, 284-90 (2015).
18. Loh, P.-R., Kichaev, G., Gazal, S., Schoech, A.P. & Price, A.L. Mixed-model association for biobank-scale datasets. *Nature genetics*, 1 (2018).
19. Ma, C., Blackwell, T., Boehnke, M., Scott, L.J. & Go, T.D.i. Recommended joint and meta-analysis strategies for case-control association testing of single low-count variants. *Genet Epidemiol* **37**, 539-50 (2013).
20. Henn, B.M. *et al.* Cryptic distant relatives are common in both isolated and cosmopolitan genetic samples. *PLoS One* **7**, e34267 (2012).
21. Delaneau, O., Zagury, J.F. & Marchini, J. Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* **10**, 5-6 (2013).
22. Browning, S.R. & Browning, B.L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* **81**, 1084-97 (2007).

23. Fuchsberger, C., Abecasis, G.R. & Hinds, D.A. minimac2: faster genotype imputation. *Bioinformatics* **31**, 782-4 (2015).
24. Zheng, X. *et al.* HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J* **14**, 192-200 (2014).
25. Banda, Y. *et al.* Characterizing Race/Ethnicity and Genetic Ancestry for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1285-95 (2015).
26. Kvale, M.N. *et al.* Genotyping Informatics and Quality Control for 100,000 Subjects in the Genetic Epidemiology Research on Adult Health and Aging (GERA) Cohort. *Genetics* **200**, 1051-60 (2015).
27. Shen, L. *et al.* The Association of Refractive Error with Glaucoma in a Multiethnic Population. *Ophthalmology* **123**, 92-101 (2016).
28. Hoffmann, T.J. *et al.* Next generation genome-wide association tool: design and coverage of a high-throughput European-optimized SNP array. *Genomics* **98**, 79-89 (2011).
29. Hoffmann, T.J. *et al.* Design and coverage of high throughput genotyping arrays optimized for individuals of East Asian, African American, and Latino race/ethnicity using imputation and a novel hybrid SNP selection algorithm. *Genomics* **98**, 422-30 (2011).
30. Delaneau, O., Marchini, J. & Zagury, J.F. A linear complexity phasing method for thousands of genomes. *Nature methods* **9**, 179-81 (2012).
31. Howie, B., Fuchsberger, C., Stephens, M., Marchini, J. & Abecasis, G.R. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature genetics* **44**, 955-9 (2012).
32. Howie, B., Marchini, J. & Stephens, M. Genotype imputation with thousands of genomes. *G3* **1**, 457-70 (2011).
33. Howie, B.N., Donnelly, P. & Marchini, J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS genetics* **5**, e1000529 (2009).
34. Marchini, J. & Howie, B. Genotype imputation for genome-wide association studies. *Nature reviews. Genetics* **11**, 499-511 (2010).
35. Chang, C.C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *GigaScience* **4**, 7 (2015).
36. Huang, L., Wang, C. & Rosenberg, N.A. The relationship between imputation error and statistical power in genetic association studies in diverse populations. *American journal of human genetics* **85**, 692-8 (2009).
37. Price, A.L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics* **38**, 904-9 (2006).
38. Verhoeven, V.J. *et al.* Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet* **45**, 314-8 (2013).
39. Riboli, E. & Kaaks, R. The EPIC Project: rationale and study design. European Prospective Investigation into Cancer and Nutrition. *Int J Epidemiol* **26 Suppl 1**, S6-14 (1997).
40. Day, N. *et al.* EPIC-Norfolk: study design and characteristics of the cohort. European Prospective Investigation of Cancer. *Br J Cancer* **80 Suppl 1**, 95-103 (1999).
41. Hayat, S.A. *et al.* Cohort profile: A prospective cohort study of objective physical and cognitive capability and visual health in an ageing population of men and women in Norfolk (EPIC-Norfolk 3). *Int J Epidemiol* **43**, 1063-72 (2014).
42. Khawaja, A.P. *et al.* The EPIC-Norfolk Eye Study: rationale, methods and a cross-sectional analysis of visual impairment in a population-based cohort. *BMJ Open* **3**(2013).

43. Delaneau, O., Marchini, J., Genomes Project, C. & Genomes Project, C. Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel. *Nat Commun* **5**, 3934 (2014).
44. Willer, C.J., Li, Y. & Abecasis, G.R. METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* **26**, 2190-1 (2010).
45. Yang, J. *et al.* Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nat Genet* **44**, 369-75, S1-3 (2012).
46. Yang, J., Lee, S.H., Goddard, M.E. & Visscher, P.M. GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* **88**, 76-82 (2011).
47. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* **42**, 565-9 (2010).
48. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, 289-300 (1995).
49. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997-1004 (1999).
50. Bulik-Sullivan, B.K. *et al.* LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* **47**, 291-5 (2015).
51. Bulik-Sullivan, B. *et al.* An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236-41 (2015).
52. Wood, A.R. *et al.* Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* **46**, 1173-86 (2014).
53. Dayem Ullah, A.Z., Lemoine, N.R. & Chelala, C. SNPnexus: a web server for functional annotation of novel and publicly known genetic variants (2012 update). *Nucleic Acids Res* **40**, W65-70 (2012).
54. McKusick, V.A. Mendelian Inheritance in Man and its online version, OMIM. *Am J Hum Genet* **80**, 588-604 (2007).
55. Buniello, A. *et al.* The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res* **47**, D1005-D1012 (2019).
56. Pruim, R.J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336-7 (2010).
57. Carnes, M.U., Allingham, R.R., Ashley-Koch, A. & Hauser, M.A. Transcriptome analysis of adult and fetal trabecular meshwork, cornea, and ciliary body tissues by RNA sequencing. *Exp Eye Res* **167**, 91-99 (2018).
58. Ratnapriya, R. *et al.* Retinal transcriptome and eQTL analyses identify genes associated with age-related macular degeneration. *Nat Genet* **51**, 606-610 (2019).
59. Finucane, H.K. *et al.* Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat Genet* **50**, 621-629 (2018).
60. Zhu, Z. *et al.* Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets. *Nat Genet* **48**, 481-7 (2016).
61. Westra, H.J. *et al.* Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet* **45**, 1238-1243 (2013).
62. Qi, T. *et al.* Identifying gene targets for brain-related traits using transcriptomic and methylomic data from blood. *Nat Commun* **9**, 2282 (2018).
63. Segre, A.V. *et al.* Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet* **6**(2010).
64. Subramanian, A. *et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-50 (2005).

65. Pybus, M. *et al.* 1000 Genomes Selection Browser 1.0: a genome browser dedicated to signatures of natural selection in modern humans. *Nucleic Acids Res* **42**, D903-9 (2014).
66. Voight, B.F., Kudaravalli, S., Wen, X. & Pritchard, J.K. A map of recent positive selection in the human genome. *PLoS Biol* **4**, e72 (2006).
67. Sabeti, P.C. *et al.* Genome-wide detection and characterization of positive selection in human populations. *Nature* **449**, 913-8 (2007).
68. Zhang, Y., Qi, G., Park, J.H. & Chatterjee, N. Estimation of complex effect-size distributions using summary-level statistics from genome-wide association studies across 32 complex traits. *Nat Genet* **50**, 1318-1326 (2018).